BRANCHING PROCESSES IN DISEASE EPIDEMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Sarabjeet Singh

May 2014

BRANCHING PROCESSES IN DISEASE EPIDEMICS

Sarabjeet Singh, Ph.D.

Cornell University 2014

Branching processes have served as a model for chemical reactions, biological growth processes and contagion (of disease, information or fads). Through this connection, these seemingly different physical processes share some common universalities that can be elucidated by analyzing the underlying branching process. In this thesis, we focus on branching processes as a model for infectious diseases spreading between individuals belonging to different populations. The distinction between populations can arise from species separation (as in the case of diseases which jump across species) or spatial separation (as in the case of disease spreading between farms, cities, urban centers, etc). A prominent example of the former is zoonoses – infectious diseases that spill from animals to humans – whose specific examples include Nipah virus, monkeypox, HIV and avian influenza. A prominent example of the latter is infectious diseases of animals such as foot and mouth disease and bovine tuberculosis that spread between farms or cattle herds. Another example of the latter is infectious diseases of humans such as H1N1 that spread from one city to another through migration of infectious hosts.

This thesis consists of three main chapters, an introduction and an appendix. The introduction gives a brief history of mathematics in modeling the spread of infectious diseases along with a detailed description of the most commonly used disease model – the Susceptible-Infectious-Recovered (SIR) model. The introduction also describes how the stochastic formulation of the model reduces to a branching process in the limit of large population which is analyzed in detail. The second chapter describes a two species model of

zoonoses with coupled SIR processes and proceeds into the calculation of statistics pertinent to cross species infection using multitype branching processes. The third chapter describes an SIR process driven by a Poisson process of infection spillovers. This is posed as a model of infectious diseases where a 'reservoir' of infection exists that infects a susceptible host population at a constant rate. The final chapter of the thesis describes a general framework of modeling infectious diseases in a network of populations using multitype branching processes.

Biographical Sketch

Sarabjeet Singh was born and brought up in New Delhi by loving parents, Ajit Singh and Inderjeet Kaur. He came from humble beginnings and was motivated by his parents to work hard right from his early years. Post high school, he went to the Indian Institute of Technology Delhi and pursued a Bachelors degree in Mechanical Engineering. Years later he would regret the decision of not pursuing Mathematics or Physics when he had the opportunity to do so. During his last year at college, he got a taste of research while working on his Bachelor's thesis and from that experience made up his mind to pursue a graduate degree. Although he started his PhD with the intention of pursuing research in rigid body mechanics, his interests changed to applied mathematics and he ended up in Mathematical Epidemiology, a field that he knew nothing about and for which his undergraduate training would not come in use. Yet he decided to jump in and embrace the uncertainty as he does so in other aspects of life. In the end he was satisfied with what he had achieved and gladly moved on to new adventures outside of academia.

To my family who has supported me all through my life, to my wife Shireen who has been an equal partner in my journey of last 5 years, and to my teachers from high school right up till graduate school who have taught me everything I know.

Acknowledgments

I am deeply indebted to my advisor Chris Myers for his guidance and support throughout my PhD. He has helped me in becoming a better researcher and most importantly a better academic writer. I would also like to acknowledge the support from the rest of AFIDD group that I was a part of: Jason Hindes, Oleg Kogan, Dave Schneider, Marshall Hayes and Drew Dolgert.

I am grateful to my parents, Ajit Singh and Inderjeet Kaur for their love, support and encouragement during my graduate studies, to my wife who has been my beacon of strength for the past 3 years and made me a better person, to my brother Manpreet Singh, my cousins Jaspreet Singh and Jasmeet Kaur with whom I share loving memories of my childhood, to my dear friends Nitin GR, Manoj Singhal and Gagandeep Singh for being there when I needed them the most.

I would like to acknowledge my high school math teacher Mr. Dilip Sethia, who taught me mathematics from 6^{th} grade to 12^{th} grade. My love for the subject started in his classroom and I shall always be grateful to him.

In the end, I would also like to thank the Science & Technology Directorate, Department of Homeland Security for providing financial support for my research via the grant HSHQDC-10-X-00138.

Table of Contents

List of Figures

CHAPTER 1

Chapter 1: Introduction

This thesis is about branching processes in mathematical epidemiology. Epidemiology is a branch of biological science that deals with the incidence, distribution, and control of disease in a population. Mathematical epidemiology is a branch of applied science that marries mathematics and statistics with the phenomenology of contagion processes. This thesis specifically deals with infectious diseases as the contagion process but the models and methods are broadly applicable to studying the propagation of rumors, fads, information and computer viruses. Branching processes are a class of stochastic mathematical models characterized by multiplicative growth starting with some initial population. Each generation produces offspring independent of other individuals in the population (see figure 1.1) and the process as a whole either grows indefinitely or dies out after some number of generations. Almost all epidemiological models that describe host to host transmission share some universal characteristics of branching processes. A disease that fails to invade a population is equivalent to a branching process that dies out. On the other hand, a disease that spreads efficiently in a population can be mapped to a growing branching process. By virtue of this universality, many structural characteristics of disease models can be unveiled by analyzing the underlying branching process.

In this chapter we provide a historical introduction of mathematics in epidemiology and proceed to one of the most commonly used model of epidemics: the Susceptible-Infected-Recovered (SIR) model. We describe both the deterministic and stochastic formulations of

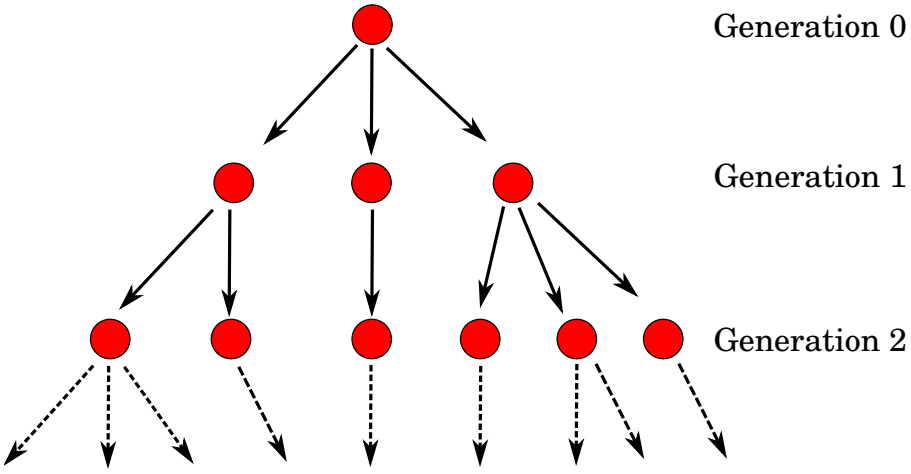


Figure 1.1: A schematic of a branching process

the model and discuss how the stochastic model reduces to a linear birth-death process which is a continuous time branching process. Further analysis of the linear birth-death process is presented as a prelude to forthcoming chapters where more complex models are introduced.

1.1 Mathematics in Epidemiology

Mathematics in epidemiology has a historical precedent. The first mathematical model in epidemiology was developed by Daniel Bernoulli in 1760s. The main objective of Bernoulli was to calculate the gain in life expectancy at birth if smallpox were to be eliminated as a cause of death [1, 2]. His model compared the change in average life expectancy if every child was inoculated against smallpox at birth. His analysis came out at a time when inoculation was being strongly debated in Europe. And though there was a controversy surrounding his arguments [3], his work was hailed as the first systematic treatment of an epidemiological problem. In another example, years later John Snow published a detailed statistical analysis of cholera in 1855 [4]. His investigation and statistical analysis of spatial

data from the 1854 cholera outbreak in London led to the identification of the contaminated water source that was the definitive cause of the outbreak [5].

Modern mathematical epidemiology has its roots in the twentieth century with the pioneering work done by Hamer, Ross, Macdonald, Kermack and McKendrick [6]. Hamer formulated a discrete time model to understand recurring epidemics of measles [7]. Ronald Ross is credited with the discovery of mosquito based transmission of malaria. He along with George Macdonald are credited with developing a mathematical model of mosquito-borne pathogen transmission [8]. Kermack and McKendrick published a series of papers on what is now referred to as the Kermack-McKendrick SIR model. They fitted their model to the 1906 Bombay plague [9] and were able to conclude the following. (1) There exists a threshold density of population which depends on the infectivity, recovery and death rates peculiar to the epidemic. No epidemic can occur if the population density is below this threshold value. (2) A small increase in infectivity rate may lead to large epidemics. (3) An epidemic, in general comes to an end before the susceptible population has been exhausted. These conclusions were derived from the theory presented in their seminal paper which we shall explore in detail in the next sections.

1.2 The Kermack-McKendrick SIR model

The SIR model is deterministic and specified through ordinary differential equations (ODE):

$$\begin{aligned} \frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1.1}$$

$$S(t) + I(t) + R(t) = N$$

The model is categorized as a compartmental model where N host individuals can be in either of the Susceptible (S), Infected/Infectious (I) or Removed/Recovered (R) compartments at any point in time. In this simple model, no distinction is made between infected and infectious hosts, i.e., it is assumed that as soon as a host is infected he/she becomes infectious and capable of infecting others. In contrast, the SEIR model includes the Exposed (E) compartment to make this distinction. The R compartment holds the individuals that are no longer infectious. Individuals may recover and develop immunity or die (and be removed) depending on the particular disease being modeled but in the context of this model those are considered equivalent. Henceforth, we shall restrict our terminology to 'Infected' and 'Recovered'. The equations 1.1 describe the rate of transitioning between the three states. β is the rate of infectious contact and γ is the rate of recovery/removal. The infectious period is defined as the average duration of time that an infected host remains infected and is given by $1/\gamma$ for the Kermack-McKendrick SIR model. This model is suitable for epidemics that sweep through the population on fast timescales on which other ecological processes such as births and deaths in the population can be ignored.

The SIR model of eq. 1.1 is an example of a fully mixed model where an infected host can infect any susceptible host in the population. This is a strong assumption regarding the host contact patterns which may not be justified for certain human diseases such as sexually transmitted diseases (STD). The assumption is justified for diseases such as measles where the predominant hosts (children) do mix strongly in localized environments (such as in schools and playgrounds). Other examples are infectious disease that involve cattle or poultry. Since farms are densely packed, the assumption of full mixing is appropriate.

From eq. 1.1, it can be verified that any small introduction of disease in the form of I_0 initial infections would grow if dI/dt is positive. This will happen when the following

condition is satisfied

$$\frac{\beta S_0}{\gamma} > 1 \tag{1.2}$$

In a population with large N and $I_0 \sim O(1)$, we can assume that $S_0 \approx N$ such that the condition 1.2 becomes

$$\frac{\beta N}{\gamma} > 1 \tag{1.3}$$

In their seminal paper, Kermack and McKendrick gave an alternate interpretation of the threshold condition as

$$N > \frac{\gamma}{\beta} \tag{1.4}$$

i.e., that the host population needs to be above a critical threshold in order for the epidemic to spread. The expression $\beta N/\gamma$ is known as the basic reproduction number and will be denoted as R_0 or α throughout this thesis. The basic reproduction number represents the average number of new infections produced by an infectious host in a fully susceptible population and determines whether any initial infection takes off or dies out:

$$\alpha \equiv R_0 = \frac{\beta N}{\gamma} \tag{1.5}$$

Figure 1.2 shows a representative numerical solution of the deterministic SIR for $\alpha > 1$.

1.2.1 Density dependent vs. frequency dependent transmission

A variation on the SIR model presented earlier is as follows:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N} SI \\ \frac{dI}{dt} &= \frac{\beta}{N} SI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1.6}$$

$$S(t) + I(t) + R(t) = N$$

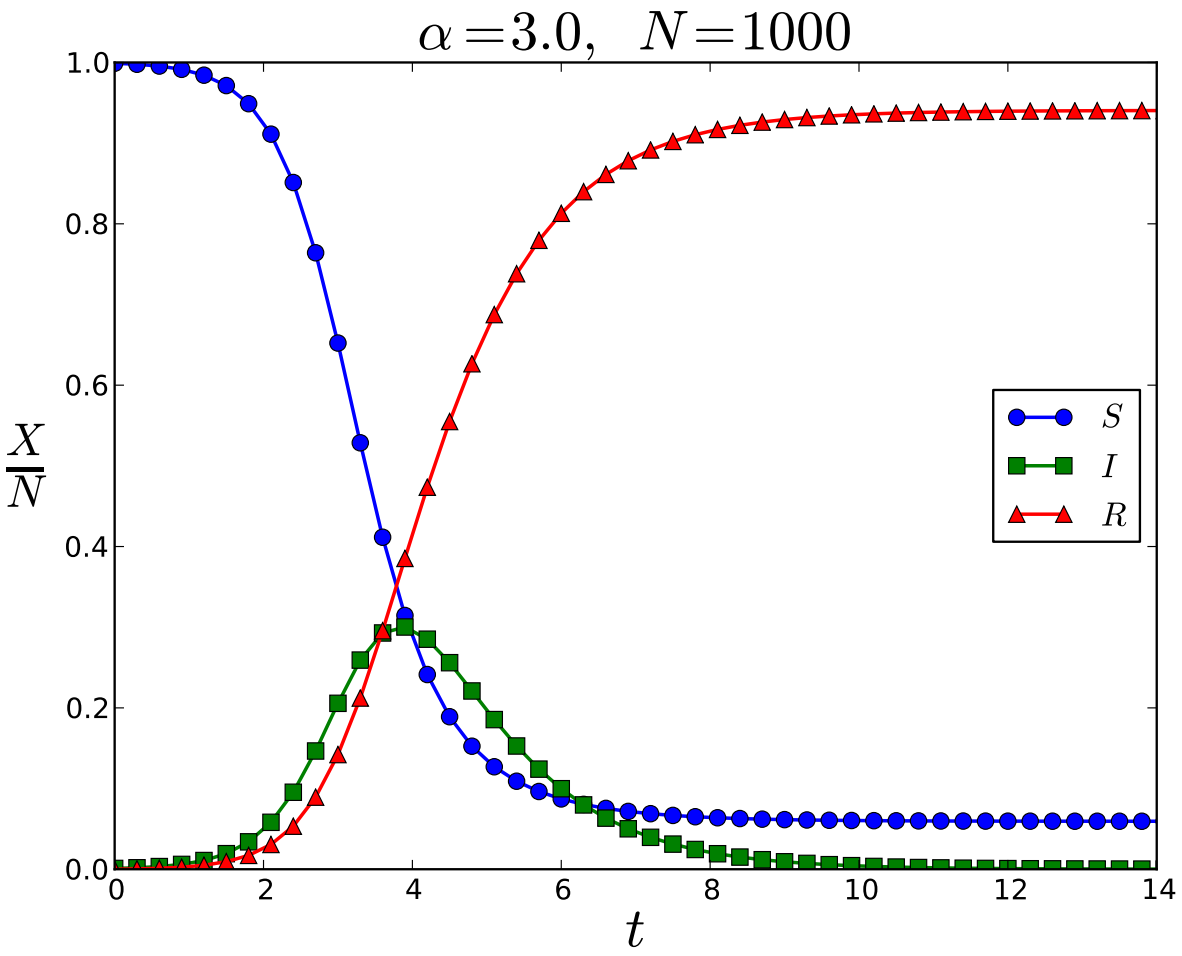


Figure 1.2: Solution of the ODE system (eq. 1.1) for $N = 10^3, \alpha = 3.0$ (see eq. 1.5 for definition of α). Time is rescaled by the rate of recovery.

In the above set of equations, the rate of infectious contact β is diluted by the population size N. While the former (eq. 1.1) is an example of density-dependent transmission, the latter (eq. 1.6) is an example of frequency-dependent transmission. In density-dependent transmission, the rate of infectious contact per infected host, βS, increases with the system size whereas in the case of frequency-dependent transmission, the rate $\beta S/N$ is unaffected by rescaling of N (because S is rescaled accordingly). In other words, the population size affects the host level transmission rate in the case of density-dependent transmission but not so in the case of frequency-dependent transmission. For eq. 1.6, the basic reproduction number is a rescaled version of the one defined for eq. 1.1, i.e.,

$$\alpha \equiv R_0 = \frac{\beta}{\gamma} \tag{1.7}$$

For an SIR model with finite N, there is absolutely no difference in the dynamics of the model. The two models differ by rescaling of β to β/N and correspondingly the definition of the basic reproduction number. As a result, the notion of a population threshold is only valid for SIR with density-dependent transmission but not so in the case of frequency-dependent transmission. However, both formulations yield the same non-dimensional forms, i.e.,

$$\begin{aligned} \frac{d\mathcal{S}}{d\tau} &= -\alpha\mathcal{S}\mathcal{I} \\ \frac{d\mathcal{I}}{d\tau} &= \alpha\mathcal{S}\mathcal{I} - \mathcal{I} \\ \frac{d\mathcal{I}}{d\tau} &= \mathcal{I} \end{aligned} \tag{1.8}$$

$$\mathcal{S}(\tau) + \mathcal{I}(\tau) + \mathcal{R}(\tau) = 1, \quad \tau = \gamma t$$

$$\mathcal{S}(\tau) = \frac{S(\tau)}{N},\ \mathcal{I}(\tau) = \frac{I(\tau)}{N},\ \mathcal{R}(\tau) = \frac{R(\tau)}{N}$$

and as a result have the same dynamics for a given α. An important distinction between the two processes becomes apparent in the limit of $N \to \infty$, in which case the model with

frequency-dependent transmission is much easier to work with. This limit when applied to the stochastic version of the model yields a continuous time branching process which will be a subject of much interest throughout this thesis. It is for this reason that we will henceforth consider SIR models with frequency-dependent transmission.

1.2.2 Final size of epidemic

Using the non-dimensional representation (eq. 1.8), we can calculate the relative final size of the epidemic, i.e., the fraction of hosts ever infected.

$$\frac{d\mathcal{S}}{d\mathcal{R}} = -\alpha S \tag{1.9}$$

The ODE can be integrated from $t = 0$ to $t = \infty$ with the initial conditions $\mathcal{S}(0) = 1 - \epsilon, \mathcal{R}(0) = 0$. In the limit of large N, $\epsilon \to 0$ and we obtain

$$1 - \mathcal{R}(\infty) = e^{-\alpha\,\mathcal{R}(\infty)} \tag{1.10}$$

For $\alpha > 1$, $\mathcal{R}(\infty)$ lies between 0 and 1. Figure 1.3 shows the numerical solution for $\mathcal{R}(\infty)$ as a function of the basic reproduction number α.

1.3 Stochastic SIR

The stochastic version of the simple SIR model is represented using the following rate equations:

$$\begin{aligned} (S, I, R) &\xrightarrow{\beta SI/N} (S-1, I+1, R) \\ (S, I, R) &\xrightarrow{\gamma I} (S, I-1, R+1) \end{aligned} \tag{1.11}$$

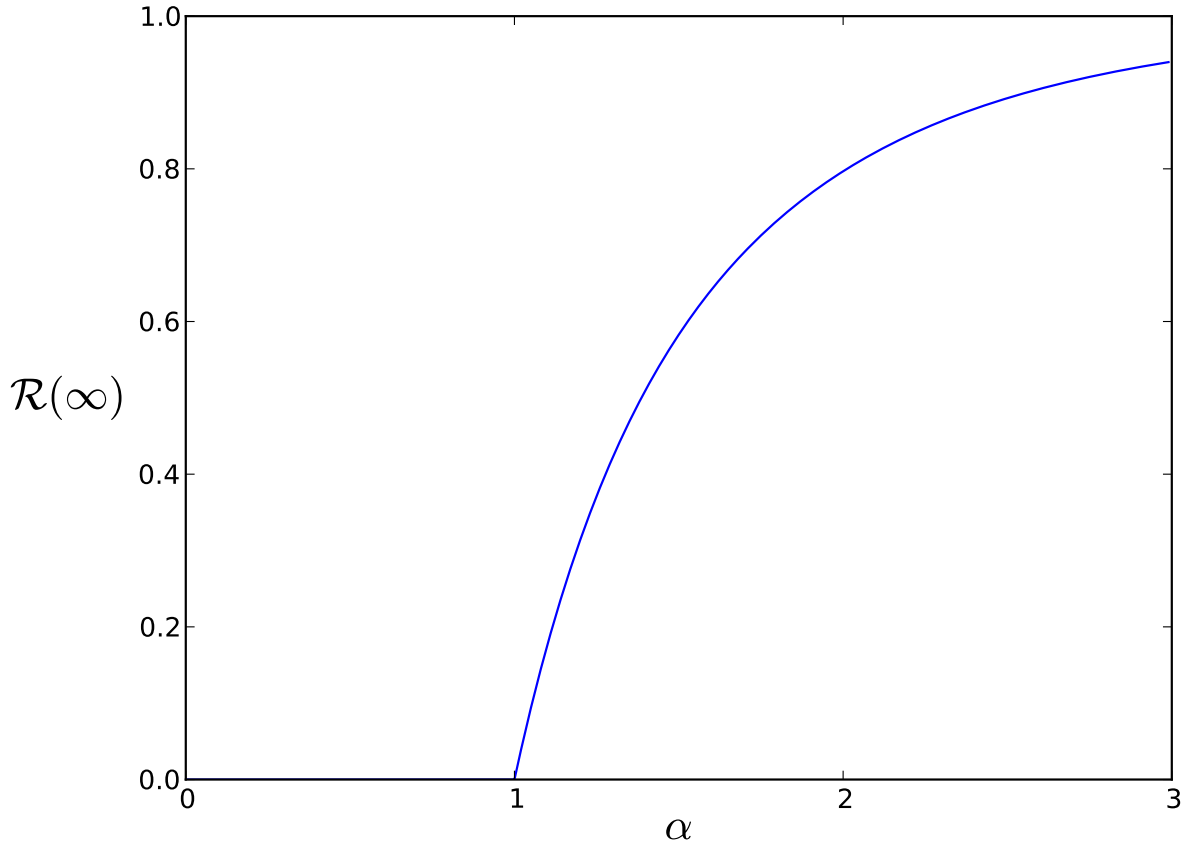


Figure 1.3: Solution of the equation 1.10 relating α and the relative final size of the epidemic.

Any single realization of the epidemic process would be henceforth termed as an *outbreak*. The process is a continuous time Markov chain and the rates represent probabilities per unit time of occurrence of a particular reaction, i.e.,

$$\mathbb{P}[S(t+dt) = s-1, I(t+dt) = i+1 | S(t) = s, I(t) = i] = \beta si/N\, dt$$
$$\mathbb{P}[I(t+dt) = i-1, R(t+dt) = r+1 | I(t) = i, R(t) = r] = \gamma i\, dt$$

In the limit of $N \to \infty$, the model reduces to a linear birth-death process [10] with the following rate equations.

$$\begin{aligned} (I, R) &\xrightarrow{\beta I} (I+1, R) \\ (I, R) &\xrightarrow{\gamma I} (I-1, R+1) \end{aligned} \tag{1.12}$$

The limit of large population ensures an infinite susceptible pool preventing any saturation in the epidemic in finite time.

1.3.1 Distribution of linear birth-death process

The description of the dynamics arising from process 1.12 can be expressed in the form of probability $Q_{m,n}(t)$ of the state variables $(I(t), R(t))$ being in the state (m, n) at time t given that the process starts with 1 infectious host at time 0. The set of all possible transitions that involve a single infectious host are: (1) production of another infectious host with rate β or (2) moving into recovered state with rate γ. The probabilities of being in the state (m, n) at time t from these one step transitions are $Q^{1\to 2}_{m,n}(t)$ and $Q^{1\to\emptyset}_{m,n}(t)$. respectively. These probabilities satisfy the Kolmogorov backward equation [11]:

$$\frac{dQ_{m,n}}{dt} = \beta Q^{1\to 2}_{m,n} + \gamma Q^{1\to\emptyset}_{m,n} - (\beta+\gamma) Q_{m,n} \tag{1.13}$$

If the single infectious host recovers before infecting other hosts then the only state that can be achieved is $(0, 1)$. Thus, $Q^{1\to\emptyset}_{m,n}(t) = \delta_{m,0}\delta_{n,1}$.

The set of backward equations 1.13 can be solved by using probability generating functions (PGF, see Appendix A). The PGF for the linear birth-death process is defined as

$$G(x, y; t) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} Q_{m,n}(t)\, x^m y^n \tag{1.14}$$

Multiplying eq. 1.13 with $x^m y^n$ and summing yields a partial differential equation (PDE) for the generating function G:

$$\frac{\partial G}{\partial t} = \beta G^2 - (\beta + \gamma)G + \gamma y \tag{1.15}$$

where we have used the fact that the PGF for $Q_{m,n}^{1\to 2}$ is $[G(x, y; t)]^2$. This is because once the single infectious host produces another infection, the branching processes emanating out of the two infectious hosts are independent and their PGF can be multiplied to yield the PGF of the original process (see Appendix A.3). The PDE 1.15 is solved with the initial condition,

$$G(x, y; 0) = x \tag{1.16}$$

which states that there is a single infected host at time 0. The PDE and its solution are provided in [11, 12] and we reproduce some of the steps of derivation here. Since the PDE involves only one partial derivative, it can be solved as an ODE:

$$\frac{dG}{dt} = \beta(G - \Lambda_0)(G - \Lambda_1) \tag{1.17}$$

where $\Lambda_0(y)$ and $\Lambda_1(y)$ are roots of the following quadratic equation such that $0 < \Lambda_0 < 1 < \Lambda_1$:

$$\alpha w^2 - (\alpha + 1)\, w + y = 0 \tag{1.18}$$

The ODE (eq. 1.17) can now be solved by separation of variables and it results in the following solution:

$$G(x, y; t) = \frac{\Lambda_0(\Lambda_1 - x) + \Lambda_1(x - \Lambda_0)e^{-\beta(\Lambda_1 - \Lambda_0)t}}{(\Lambda_1 - x) + (x - \Lambda_0)e^{-\beta(\Lambda_1 - \Lambda_0)t}} \tag{1.19}$$

The PGFs for the marginal distributions of $I(t)$ and $R(t)$ are $G(x,1;t)$ and $G(1,y;t)$ respectively. The probability of extinction as a function of time is $G(0,1;t)$

$$Q_0(t) = \begin{cases} \dfrac{1-e^{\gamma(\alpha-1)t}}{1-\alpha e^{\gamma(\alpha-1)t}} & \alpha \neq 1, \\ \dfrac{\beta t}{1+\beta t} & \alpha = 1. \end{cases} \tag{1.20}$$

where $\alpha = \beta/\gamma$ is the basic reproduction number. From eq. 1.20, we can estimate the overall probability of extinction.

$$Q_0(\infty) = \begin{cases} 1 & \text{if } \alpha \leq 1 \\ \dfrac{1}{\alpha} & \text{if } \alpha > 1. \end{cases} \tag{1.21}$$

Thus, $\alpha = 1$ is a critical point in the process. For $\alpha \leq 1$, the process becomes extinct with probability 1 whereas for $\alpha > 1$, a large outbreak occurs with probability $1 - 1/\alpha$. A large outbreak is characterized by an infinite number of infected hosts. For the original SIR process, this corresponds to a characteristic size $\mathcal{O}(N)$ in the limit of $N \to \infty$. Thus, an outbreak is defined to be small if its relative size is 0 in the limit of large N, i.e.,

$$\lim_{N\to\infty} \frac{R(\infty)}{N} \to 0$$

whereas an outbreak is large if its relative size is positive

$$0 < \lim_{N\to\infty} \frac{R(\infty)}{N} < 1$$

Moreover, the relative size of the large outbreak approaches the deterministic value (eq. 1.10) in the same limit, i.e.,

$$\lim_{N\to\infty} \frac{R(\infty)}{N} \to \mathcal{R}(\infty)$$

The convergence is in distribution and occurs via a central limit type argument [13, 14]. Note that extinction in the case of $\alpha > 1$ is only observed in the stochastic model but not the deterministic one.

1.3.2 Distribution of outbreak sizes

From the PGF in eq. 1.19, we calculate the distribution of outbreak sizes $R(\infty)$. Let $\mathbb{P}[R(\infty) = n] \equiv P(n)$ be generated by the PGF $H(y)$:

$$H(y) \equiv \sum_{n=0}^{\infty} P(n)\, x^n = \frac{1 + \alpha - \sqrt{(1+\alpha)^2 - 4\alpha y}}{2\alpha} \tag{1.22}$$

The average outbreak size when $\alpha < 1$ is obtained as $H'(1)$:

$$H'(1) = \frac{1}{1-\alpha} \tag{1.23}$$

which diverges as $\alpha \to 1$ as expected. For $\alpha > 1$, the probability distribution contained in $H(y)$ is defective. Thus, conditioning on the occurrence of a small outbreak, the average outbreak size is calculated as

$$\begin{aligned} \langle n \rangle_s &= \frac{H'(1)}{H(1)} \\ &= \begin{cases} \dfrac{1}{1-\alpha} & \alpha < 1, \\ \dfrac{\alpha}{\alpha - 1} & \alpha > 1. \end{cases} \end{aligned} \tag{1.24}$$

The distribution $P(n)$ can be sampled numerically from the PGF using the Cauchy integral formula (A.4). Near the critical point at $\alpha = 1$, the asymptotic form of $P(n)$ can be obtained by singularity analysis of the generating function (see Appendix D.3):

$$P(n) \sim \zeta^{-n}\, n^{-3/2}, \quad n \to \infty \tag{1.25}$$

where

$$\zeta = \frac{4\alpha}{(1+\alpha)^2}$$

This is a well-known result in the literature [15]. The scaling for $P(n)$ becomes a pure power-law when $\alpha = 1$ which is typical of a physical phenomenon undergoing a continuous phase transition.

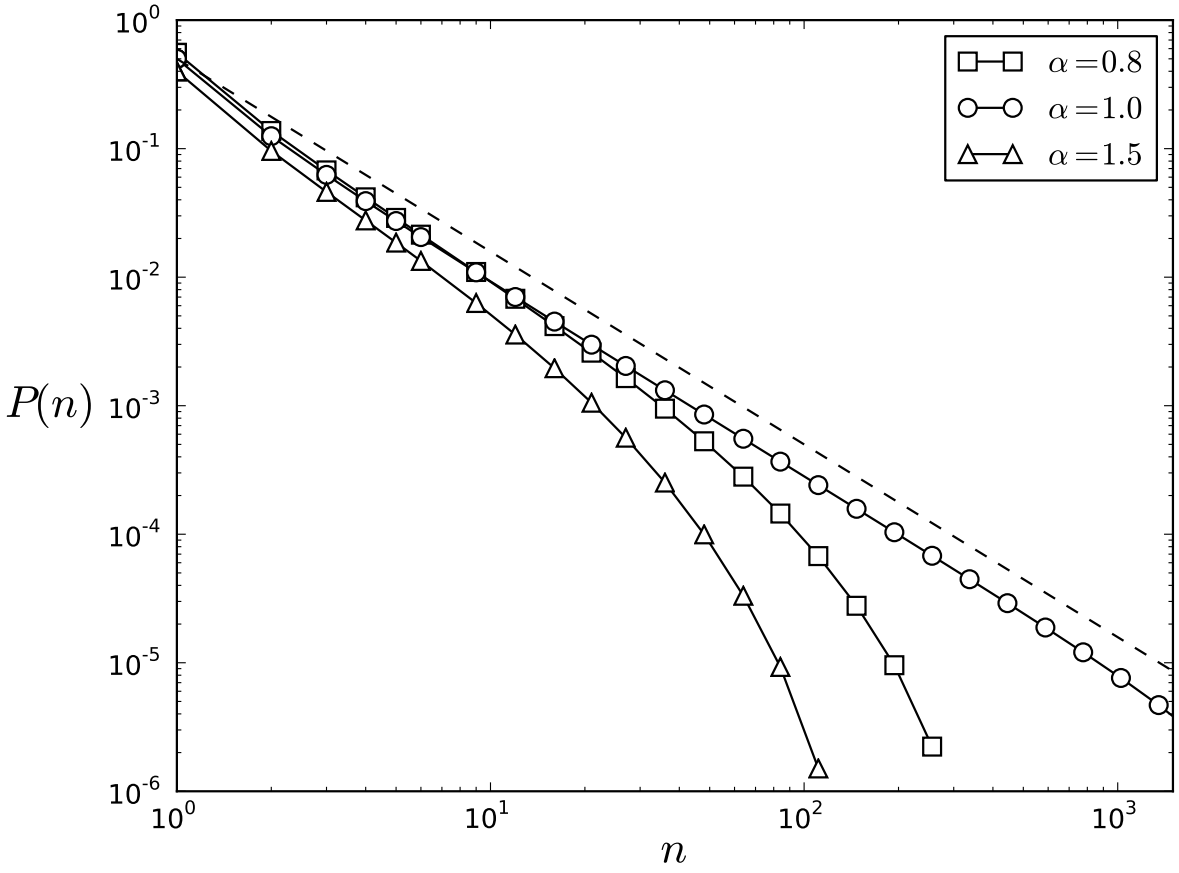


Figure 1.4: The distribution of outbreak sizes in a linear birth-death process, sampled from the PGF $H(y)$ (eq. 1.22) for different values of α. The dashed line is a pure power-law with slope -3/2.

1.3.3 Interpretation as Galton-Watson branching process

The distribution of outbreak sizes in the linear birth-death process can also be calculated by interpreting the process as a Galton-Watson branching process (see Appendix B). The infections caused by a direct transmission constitute the offspring of a given host. The generation to which an individual belongs is measured as the tree distance from the first infected host. In this interpretation, the offspring distribution is calculated as follows. Each individual 'produces' offspring at a constant rate β until they are infectious. The infectious period is an exponentially distributed random variable with rate γ. The PGF of the offspring distribution is given by

$$F(x) = \int_0^\infty e^{\beta(x-1)t}\, \gamma e^{-\gamma t}\, dt = \frac{1}{1-\alpha(x-1)} \tag{1.26}$$

The PGF $F(x)$ indicates that the offspring distribution is geometrically distributed. Using eq. B.6, the distribution of the total population in the branching process is given by the following equation.

$$H(y) = yF(H(y)) \tag{1.27}$$

whose solution is obtained by inserting the expression for $F(x)$ from eq. 1.26. This leads to the same PGF as was obtained from the distribution of the linear birth-death process (eq. 1.22).

1.3.4 Finite size scaling at $\alpha = 1$

While the average size of large outbreak is $\mathcal{O}(N)$ and that of small outbreaks is $\mathcal{O}(1)$, outbreaks occurring in a critical process ($\alpha = 1$) show a sub-linear scaling of $\mathcal{O}(N^{1/3})$ [16]. This can be derived in the following manner using methods outlined in [15, 16]. In a finite

size system, the power-law scaling of outbreak size will extend upto the scale of the maximal outbreak size. No outbreaks can exceed this scale in size due to finite size of the system. Let this scale be denoted by M. From eq. 1.25, we can calculate the scale of the average outbreak size as,

$$\langle n \rangle \sim \sum_{n=1}^{M} n \cdot n^{-3/2} \sim \sqrt{M} \tag{1.28}$$

Due to finite N, the effective infection rate becomes $\alpha_\star = 1 - M/N$ when the outbreak size reaches the scale M. Using eq. 1.23, a second estimate of the average outbreak size is

$$\langle n \rangle \sim \frac{N}{M} \tag{1.29}$$

Equating the two estimates leads to the following:

$$\langle n \rangle \sim N^{1/3}, \quad M \sim N^{2/3} \tag{1.30}$$

The scaling of the average duration can be calculated in a similar way. From eq. 1.20, the distribution of the outbreak duration T for the linear birth-death process when $\alpha < 1$ is given by

$$\mathbb{P}[T \leq t] = Q_0(t) = \frac{1 - e^{\gamma(\alpha-1)t}}{1 - \alpha e^{\gamma(\alpha-1)t}} \tag{1.31}$$

From the above distribution, the average outbreak duration can be calculated as

$$\langle t \rangle = \frac{1}{\beta} \log \frac{1}{1-\alpha} \tag{1.32}$$

When N is finite, using the effective $\alpha_\star$ and M, we obtain

$$\langle t \rangle \sim \log N \tag{1.33}$$

This is another well-known result in the literature [16]. In the coming chapters, we shall investigate among other things, the finite size scaling laws for processes with more complexity.

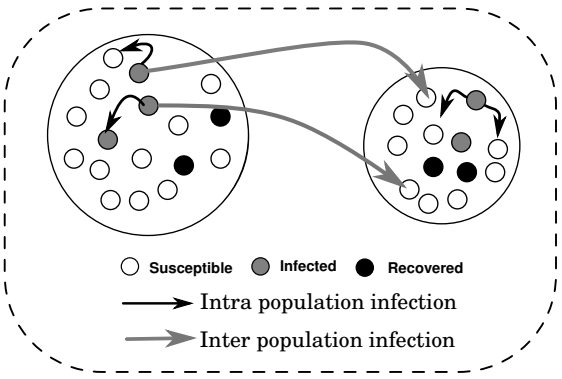


Figure 1.5: A schematic of the simplest metapopulation model with two nodes and directional coupling from node 1 to 2.

1.4 Metapopulation models

The simple SIR model can be extended to more than one host population. In this extension, infectious hosts in one population can infect susceptible hosts in a different population. The distinction between populations can arise from species separation (as in the case of diseases which jump across species) or spatial separation (as in the case of disease spreading between farms, cities, countries, etc). As an illustration, figure 1.5 shows the simplest metapopulation model with two nodes. Infected hosts in node 1 can infect susceptible hosts in both nodes 1 and 2. Let N_1 and N_2 be the populations in the two nodes and $(S_\star, I_\star, R_\star)$ be the respective state variables. The rate equations for this model are shown below.

$$\begin{aligned}
(S_1, I_1, R_1, S_2, I_2, R_2) &\xrightarrow{\beta_1 S_1 I_1/N_1} (S_1 - 1, I_1 + 1, R_1, S_2, I_2, R_2) \\
(S_1, I_1, R_1, S_2, I_2, R_2) &\xrightarrow{\gamma_1 I_1} (S_1, I_1 - 1, R_1 + 1, S_2, I_2, R_2) \\
(S_1, I_1, R_1, S_2, I_2, R_2) &\xrightarrow[+\rho_{12}\beta_1 S_2 I_1/N_2]{\beta_2 S_2 I_2/N_2} (S_1, I_1, R_1, S_2 - 1, I_2 + 1, R_2) \\
(S_1, I_1, R_1, S_2, I_2, R_2) &\xrightarrow{\gamma_2 I_2} (S_1, I_1, R_1, S_2, I_2 - 1, R_2 + 1)
\end{aligned} \tag{1.34}$$

ρ_{12} is the inter-node coupling in the model from node 1 to node 2. It is defined as the ratio of the rates of infectious contact for inter-node and intra-node transmission.

In models where the two populations represent two different species colocated spatially, the difference in the rates of infectious contact arises simply from the difference in the ecology of interactions. For example, a disease that infects both cattle and humans will have 3 different rates of contact: animal-to-animal, animal-to-human and human-to-human. The rates are specified as distinct because animals and humans interact differently. The rate of animal-to-human contact can be specified through the coupling ρ_{12}. The detailed analysis of this two-node example is the subject of chapter 3. On the other hand, in models of disease spread with spatially separated hosts, the two populations can represent different farms, cities or urban centers. In this case, the coupling ρ_{12} is the effective rate of inter-node transmission which is based on the migration of infected hosts. A paper by Frank Ball [17] describes how rates of migration can be used to calculate an effective coupling constant.

The joint distribution of outbreak sizes in the limit of $N_1, N_2 \to \infty$ are calculated using the analogy with multitype Galton-Watson processes [11]. Further, the relative size of large outbreaks can also be calculated analytically in a metapopulation model [18, 19]. The details of these methods are provided in Chapter 3.

1.5 Discussion

Mathematical epidemiology is an important tool that has been used to advise public health policy and plan control interventions in the event of disease outbreaks. The field's literature is filled with both theoretical and data-driven research that has emerged from collaborations between ecologists, biologists, mathematicians and physicists. On several occasions [20, 21], mathematical epidemiology has played a central role in making forecasts and evaluating control strategies during an ongoing disease outbreak. In other instances it has helped in explaining patterns of recurrent outbreaks [9, 22, 23].

The simple SIR model has been a cornerstone in mathematical epidemiology for which analytical results have been derived in the limit of large N. In this chapter we have presented a summary of results that are well known for the stochastic SIR model. In the next chapter, we describe the statistics of an SIR model coupled with an external force of infection.

CHAPTER 2

Chapter 2: Outbreak statistics and scaling laws for externally driven epidemics

ABSTRACT [1]

Power-law scalings are ubiquitous to physical phenomena undergoing a continuous phase transition. The classic Susceptible-Infectious-Recovered (SIR) model of epidemics is one such example where the scaling behavior near a critical point has been studied extensively. In this system the distribution of outbreak sizes scales as $P(n) \sim n^{-3/2}$ at the critical point as the system size N becomes infinite. The finite-size scaling laws for the outbreak size and duration are also well understood and characterized. In this work, we report scaling laws for a model with SIR structure coupled with a constant force of infection per susceptible, akin to a 'reservoir forcing'. We find that the statistics of outbreaks in this system are fundamentally different than those in a simple SIR model. Instead of fixed exponents, all scaling laws exhibit tunable exponents parameterized by the dimensionless rate of external forcing. As the external driving rate approaches a critical value, the scale of the average outbreak size converges to that of the maximal size, and above the critical point, the scaling laws bifurcate into two regimes. Whereas a simple SIR process can only exhibit outbreaks

[1] This chapter and the material in Appendix C have been accepted for publication in *Physical Review E* with co-author Christopher R. Myers.

of size $\mathcal{O}(N^{1/3})$ and $\mathcal{O}(N)$ depending on whether the system is at or above the epidemic threshold, a driven SIR process can exhibit a richer spectrum of outbreak sizes that scale as $\mathcal{O}(N^{\xi})$ where $\xi \in (0,1]\backslash\{2/3\}$ and $\mathcal{O}((N/\log N)^{2/3})$ at the multi-critical point.

2.1 Introduction

Epidemic models have proven to be extremely useful in understanding the spread of infectious diseases, rumors, computer viruses and fads [16, 24]. These models constitute a broader category of models describing physical processes that exhibit a second-order phase transition at a critical threshold [25]. As is characteristic of such transitions, epidemic models exhibit power-law scaling in various statistics characterizing infectious outbreaks at the critical threshold. The classic Susceptible-Infectious-Recovered (SIR) model (eq. 2.1) has been the most widely studied epidemic model in the literature [16, 26]. In this model, a small number of infected hosts start an 'outbreak' in a susceptible pool. An infected host continues to infect susceptible hosts before becoming recovered. The model can be specified using the following rate equations where the rates represent probabilities per unit time of each 'reaction' taking place:

$$\begin{aligned}(S, I, R) &\xrightarrow{\alpha SI/N} (S-1, I+1, R)\\ (S, I, R) &\xrightarrow{\quad I \quad} (S, I-1, R+1)\end{aligned} \tag{2.1}$$

Note that the rates reported in eq. 2.1 are rescaled by the rate of recovery, without loss of generality.

The SIR model has an epidemic threshold ($\alpha = 1$ in our case), below which all outbreaks are small (with size $o(N)$) and above which some outbreaks are large (with size $\mathcal{O}(N)$) [26]. At the critical threshold, the distribution of outbreak sizes shows the universal scaling of

$P(n) \sim n^{-3/2}$ which is invariant to changes in the microscopic details of the model [15, 16]. The size of the average and the maximal outbreaks scale as $N^{1/3}$ and $N^{2/3}$ at the critical point, respectively [16]. Extensions of the SIR that include multiple stages exhibit the scaling $P(n) \sim n^{-(1+2^{-p})}$ for outbreak sizes where p is the number of stages that an infected host crosses before being recovered [15]. The scaling exponents for the average and the maximal outbreak sizes in this multi-stage SIR are functions of both p and 2^{-p}, introducing a discrete degree of variability in the scaling depending on the number of stages involved.

A different extension to the simple SIR includes an external force of infection action on each susceptible:

$$\begin{aligned} (S, I, R) &\xrightarrow{\alpha S(I+\nu)/N} (S-1, I+1, R) \\ (S, I, R) &\xrightarrow{\quad I \quad} (S, I-1, R+1) \end{aligned} \tag{2.2}$$

In this system, each susceptible experiences an additional force of infection with rate $\alpha\nu/N$ where ν is a dimensionless parameter reflecting the external driving rate. Such a model describes infection dynamics where a pathogen that is sustained in a reservoir repeatedly jumps to susceptible hosts [27, 28], as might be applicable to the study of cross-species infections such as zoonotic diseases that jump from animals to humans. By construction, the model allows for re-introduction of infection after an outbreak has died out so long as there remain any susceptible host. In the subcritical case ($\alpha < 1$), the process alternates between periods of highly stochastic externally forced outbreaks and periods of no activity (see figure 2.1a). While the dynamics of this type of model have been examined previously, the calculation of the distributions of outbreak sizes and durations has surprisingly received no attention. These statistics are important for several reasons. From a theoretical perspective, we demonstrate here that outbreak statistics are qualitatively different for the externally driven system than for the simple SIR. From a practical point of

view, reservoir-driven zoonotic outbreaks are known to be sporadic [27], and time series data for outbreaks exhibit active and non-active phases over long periods of time. The information on statistics of individual outbreaks allows one to assess from given data whether the rate $\nu\alpha$ of external introduction and the rate α of infectious contact are constant over a given period of time or varying from outbreak to outbreak. In this work, we solve for the distribution of outbreak sizes $P(n)$ for this system in the limit of $N \to \infty$. From the analytical distribution, we distill scaling laws for all quantities of interest for both infinite and finite population systems at the epidemic threshold of $\alpha = 1$ with varying values of ν. In previous related work, we have calculated similar sorts of outbreak statistics for a different type of externally driven SIR system [29], arising from the coupling of an epidemic outbreak across two populations (e.g., animals and humans). The case of constant external forcing considered here would be more applicable to situations where infection is at an endemic equilibrium in the reservoir.

2.2 Infinite population

In the limit of infinite system size, the simple SIR process converges in distribution to a linear birth-death (BD) process whose analysis has provided crucial insights in to the full nonlinear process. Similarly, the distribution of the driven SIR process converges to a linear birth-death-immigration (BDI) process (eq. 2.3) as $N \to \infty$, which has been analyzed extensively in the literature [12, 30]:

$$\begin{aligned} (I, R) &\xrightarrow{\alpha(\nu + I)} (I + 1, R) \\ (I, R) &\xrightarrow{I} (I - 1, R + 1) \end{aligned} \tag{2.3}$$

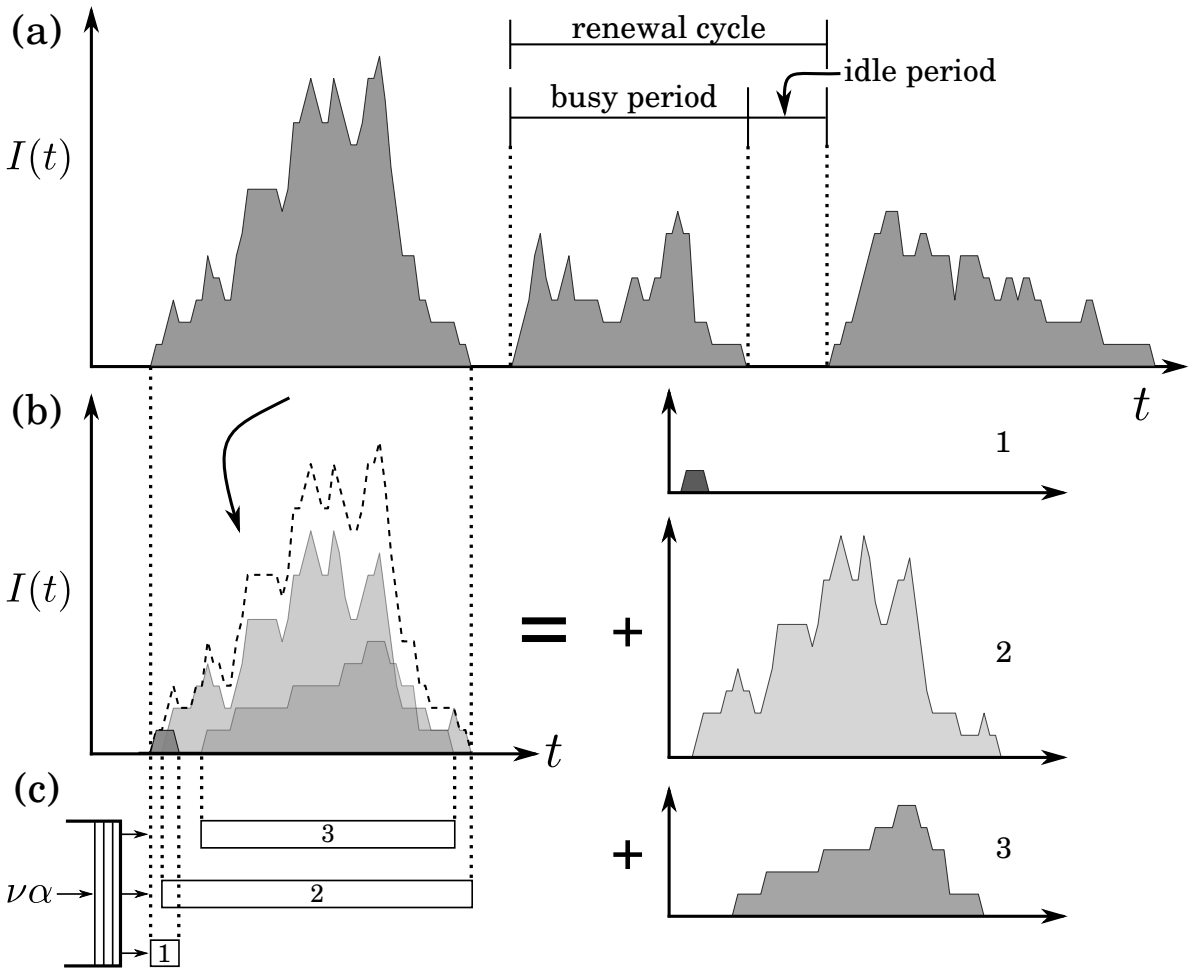


Figure 2.1: (a) A realization of the externally driven SIR process with 3 outbreaks. (b) The decomposition of an outbreak into its constituent micro-outbreaks. Each micro-outbreak is associated with a single imported infection. (c) The $M/G/\infty$ queue where a micro-outbreak is analogous to a customer being serviced at a station. In this realization, the busy period starts when the first customer enters service and ends when the second customer leaves service. The number of customers served during a busy period corresponds to the number of micro-outbreaks that constitute an outbreak. The statistics of the composite outbreaks depends strongly on the driving rate ν, which is our primary focus here.

Note that the BDI process with $\nu = 0$ is identical to the BD process. Of particular interest here are the sub-critical and critical cases ($\alpha \leq 1$) where outbreaks occur sporadically from imported infections that arrive with rate $\nu\alpha$ but go extinct with probability 1. The time-dependent distribution of the number of infected hosts in the strictly sub-critical case ($\alpha < 1$) which starts with no infection is given by a negative binomial distribution [12],

$$\mathbb{P}\big[I(t) = n\big] = \binom{n+\nu-1}{n}(1-\alpha)^{\nu}\alpha^{n}\frac{\big[1-e^{(\alpha-1)t}\big]^{n}}{\big[1-\alpha e^{(\alpha-1)t}\big]^{n+\nu}} \tag{2.4}$$

which is succinctly expressed using a probability generating function (PGF)

$$\begin{aligned} A(x;t) &= \sum_{n=0}^{\infty}\mathbb{P}\big[I(t) = n\big]x^{n} \\ &= \left[\frac{1-\alpha}{1-\alpha x + \alpha e^{(\alpha-1)t}(x-1)}\right]^{\nu} \end{aligned} \tag{2.5}$$

Due to an infinite susceptible pool and repeated introductions, the epidemic never goes extinct in the sub-critical process and the number of currently infectious hosts converges to a limiting distribution as $t \to \infty$ [12],

$$A(x;\infty) = \sum_{n=0}^{\infty}\mathbb{P}\big[I(\infty) = n\big]x^{n} = \left(\frac{1-\alpha}{1-\alpha x}\right)^{\nu} \tag{2.6}$$

The limiting sub-critical BDI process can be interpreted as a renewal process where one idle period ($I = 0$) and one busy period ($I > 0$) together form a renewal cycle [30]. We shall define an outbreak in the BDI process to be synonymous with the busy period of the renewal cycle. To obtain the distribution of outbreak sizes, we first draw the analogy between the BDI process and the $M/G/\infty$ queue, as has been reported in literature [30, 31]. The notation $M/G/\infty$ describes a queueing process where customers arrive at an infinite server station according to a Poisson process and enter into service immediately. The service time at a server has a general distribution that is specified. In the notation, M stands for Markovian arrival process, G stands for the general service time distribution and

∞ stands for the infinite number of servers [32]. The busy period for the queue is defined as the time period when at least one customer is still in service.

Each imported infection can be imagined to be an arrival in the infinite server queue. The service through a single server is then analogous to a 'micro-outbreak' in the BDI process, i.e., the chain of infections originating from a single imported infection. A micro-outbreak is then mathematically equivalent to a BD process (eq. 2.3 with $\nu = 0$) with a single infectious host at the beginning. Thus, the distribution of service times at a single server in the queue system is the same as the distribution of outbreak durations in a BD process, whose closed form solution is available [12]. Finally, the busy period of a BDI process is mathematically equivalent to the busy period of an $M/G/\infty$ queue whose statistics can be calculated using established methods in queueing theory [33]. The intuition for infinite servers comes from the fact that because outbreaks are occurring in an infinite susceptible pool, there is no constraint on how many individual micro-outbreaks can be initiated on overlapping timescales. See figure 2.1 for illustration of the preceding concepts. Before embarking on new calculations, we first report some results from the literature that we can make use of. For instance, from eq. 2.6 the probability that the limiting BDI process is in the idle state ($I = 0$) is given by

$$q_0 = (1 - \alpha)^{\nu} \tag{2.7}$$

This is equal to the probability that the equivalent $M/G/\infty$ queue is in the idle state. From queueing theory [34], we know

$$\text{Mean \# customers served in busy period} = 1/q_0 \tag{2.8}$$

The number of customers served during the busy period of the queue correspond to the number of imported infections in a single outbreak. The total outbreak size can be obtained by integrating over all the micro-outbreaks emanating from imported infections.

Heuristically, we known that the average outbreak size for the BD process is given by $(1-\alpha)^{-1}$. Using this result, we can guess that the average outbreak size for the BDI process would scale as

$$\langle n \rangle \sim \frac{1}{q_0(1-\alpha)} = (1-\alpha)^{-(1+\nu)} \tag{2.9}$$

Note that the probability of the process being in the idle state (eq. 2.7) is always greater than 0 as long as $\alpha < 1$ and ν is finite. Thus, a sub-critical BDI process can never be driven into a perpetual busy period, and accordingly there does not exist any critical driving rate.

The average duration of the outbreak can be derived using the theory of renewal processes (see [30] for a more rigorous derivation). Arrivals in the analogous $M/G/\infty$ queue form a Poisson process with rate $\nu\alpha$. A busy period begins when an arrival takes place at the end of an idle period. Thus, the renewal cycle (idle period + busy period) is a thinned Poisson process which occurs with rate $\nu\alpha q_0$ (the original rate multiplied by the probability that the arrival occurs when the cycle is in the idle state). The average duration of a renewal cycle is $1/(\nu\alpha q_0)$ and the average duration of a busy period is a fraction $1-q_0$ of the cycle duration. Combining these results, the average duration of an outbreak in the BDI process is given by

$$\langle t \rangle = \frac{1-q_0}{\nu\alpha q_0} = \frac{(1-\alpha)^{-\nu}-1}{\nu\alpha} \tag{2.10}$$

The critical BDI process ($\alpha = 1$) is an interesting analog to the critical BD process for which some results can be derived using the PGF in equation 2.5. For instance, the distribution of the number of infectious hosts as a function of time (with no infection at time 0) is generated by

$$K(x,t) = [1-(x-1)t]^{-\nu} \tag{2.11a}$$

which does not have a steady state solution. The average number of infectious hosts grows linearly with time

$$\langle I(t) \rangle = \nu t \tag{2.11b}$$

and the probability of the process being in the idle state decays with time.

$$p_0(t) = (1+t)^{-\nu} \tag{2.11c}$$

As expected, the average outbreak size (eq. 2.9) and duration (eq. 2.10) diverge at the critical threshold.

The existing generating functions for the BDI process (eq. 2.5 and 2.11a) do not describe the busy period of the process in isolation. One can only query the distribution of the number of infected (or recovered) hosts at time t without conditioning on whether the process is busy or idle and without any knowledge of how many outbreaks have occurred before t. For calculating the statistics of a single outbreak, integrating the time-dependent generating functions unconditionally would be incorrect. Instead, one must integrate over the duration of a single outbreak, which corresponds to the busy period of the analogous $M/G/\infty$ queue. The calculation for the number of customers served in the busy period exists for the $M/G/\infty$ queue [33] that we shall adopt for our purpose. The calculation presented here is done for arbitrary values of α and ν assuming that the outbreak sizes are finite. The first ingredient in this calculation are the statistics of a BD process, which are summarized in the following PGF. Let

$$F(x,y;t) = \sum_{m,n} \mathbb{P}\big[I(t)=m, R(t)=n\big]\, x^m y^n \tag{2.12}$$

be the PGF for the joint distribution of infectious and removed hosts in a BD process with

birth rate α and death rate set to 1 that starts with one infectious individual at time 0. From [12] the exact solution of the PGF is given by

$$F(x,y;t) = \frac{\Lambda_0(\Lambda_1 - x) + \Lambda_1(x - \Lambda_0)e^{-\alpha(\Lambda_1-\Lambda_0)t}}{(\Lambda_1 - x) + (x - \Lambda_0)e^{-\alpha(\Lambda_1-\Lambda_0)t}} \tag{2.13}$$

where $\Lambda_0(y)$ and $\Lambda_1(y)$ are roots of the following quadratic equation such that $0 < \Lambda_0 < 1 < \Lambda_1$.

$$\alpha w^2 - (\alpha + 1)\, w + y = 0 \tag{2.14}$$

The joint distribution of the duration T and the size $R(T)$ of an outbreak can be summarized using $F(0,y;t)$, i.e.,

$$F(0,y;t) = \sum_{n\geq 1} \mathbb{P}\big[T \leq t, R(T) = n\big]\, y^n \tag{2.15}$$

The trick that yields the desired result is to use the PGF $F(0,y;t)$ in place of the service time distribution for calculating the number of customers served in a busy period of $M/G/\infty$ queue (see Appendix C.1 for details). The intuition comes from the fact that the outbreak duration and size are correlated random variables, and integrating the joint distribution preserves the correlation. Once we substitute $F(0,y;t)$ and simplify the integration, we obtain the following PGF (eq. 2.16a) for the joint distribution of the number of imported infections and outbreak size during the busy period of the BDI process.

$$G(x,y) = 1 - \frac{1}{\nu} \frac{\Lambda_1\, z^a\, (1-z)^b}{\int_z^1 r^{a-1}(1-r)^{b-1} dr} \tag{2.16a}$$

where

$$z = 1 - \frac{\Lambda_0}{\Lambda_1}, \quad a = 1 - \nu x, \quad b = \nu\left(\frac{1 - \Lambda_0 x}{\Lambda_1 - \Lambda_0}\right) \tag{2.16b}$$

$$\Lambda_0, \Lambda_1 = \frac{(\alpha+1) \mp \sqrt{(\alpha+1)^2 - 4\alpha y}}{2\alpha} \tag{2.16c}$$

The marginal distribution of outbreak sizes is generated by $G(1, y)$. Let $H(y)$ be the PGF for the marginal distribution at the critical threshold $\alpha = 1$. This simplifies some of the terms in the PGF:

$$H(y) = 1 - \frac{1}{\nu} \frac{\Lambda_1 \, z^a \, (1-z)^b}{\int_z^1 r^{a-1}(1-r)^{b-1} dr} \tag{2.17a}$$

where

$$z = 1 - \frac{\Lambda_0}{\Lambda_1}, \quad a = 1 - \nu, \quad b = \frac{\nu}{2} \tag{2.17b}$$

$$\Lambda_0, \Lambda_1 = 1 \mp \sqrt{1-y} \tag{2.17c}$$

The integral in the denominator of eq. 2.17a can be solved explicitly for $\nu \in \mathbb{Z}_{>0}$. For an arbitrary ν, the integral can be represented as the difference between the Beta function $\mathrm{B}(a, b)$ and incomplete Beta function $\mathrm{B}(a, b; z)$. Let this integral be denoted by $\mathrm{J}(a, b; z)$.

$$\mathrm{J}(a, b; z) = \mathrm{B}(a, b) - \mathrm{B}(a, b; z) \tag{2.18}$$

The asymptotic form for $P(n)$ – the probability of having an outbreak of size n – can be obtained by the singularity expansion of the PGF $H(y)$ around $y = 1$ (or $z = 0$). For $\nu < 1$ (which implies $a, b > 0$), the incomplete Beta function can be approximated by

$$\mathrm{B}(a, b; z) \sim \frac{z^a (1-z)^b}{a} \tag{2.19}$$

in the limit of $z \to 0$ [35]. The PGF $H(y)$ simplifies as,

$$\begin{aligned} H(y) &\sim 1 - \frac{1}{\nu \mathrm{B}(a,b)} \frac{\Lambda_1 \, z^a \, (1-z)^b}{\left(1 - \dfrac{z^a (1-z)^b}{a \mathrm{B}(a,b)}\right)} \\ &= 1 - \frac{1}{\nu \mathrm{B}(a,b)} \Lambda_1 \, z^a \, (1-z)^b \left[1 + \frac{z^a (1-z)^b}{a \mathrm{B}(a,b)} + \cdots\right] \\ &= 1 - \frac{2^{1-\nu}(1-y)^{(1-\nu)/2}(y)^{\nu/2}}{\nu \mathrm{B}(a,b)} + \cdots \end{aligned} \tag{2.20}$$

where the simplification in the last step follows from substituting for z from eq. 2.17b. From the leading order term $(1-y)^{(1-\nu)/2}$, we can assert the following asymptotic form for $P(n)$ as described in [36]

$$P(n) \sim n^{-(3-\nu)/2} \tag{2.21}$$

The scaling law is verified in figure 2.2. As expected, the power-law becomes more and more flat with increasing ν. The pronounced bump in the simulations is a finite size effect due to the clustering of outbreaks that would have continued to exhibit the power-law scaling if the system size was infinite [37]. It can be verified that all moments of the distribution diverge for any value of $\nu \in [0, 1)$.

At $\nu = 1$, the PGF $H(y)$ in eq. 2.17a can be simplified further into a closed form solution,

$$H(y) = 1 + \frac{\sqrt{y}}{\log\left(\dfrac{\sqrt{1-y}}{1+\sqrt{y}}\right)} \tag{2.22}$$

whose singularity analysis around $y = 1$ yields the following asymptotic form for $P(n)$

$$P(n) \sim \frac{1}{n \log^2 4n}\left[\frac{1}{2} - \frac{\gamma}{\log 4n} + \mathcal{O}\left(\frac{1}{\log^2 4n}\right)\right] \tag{2.23}$$

where $\gamma = 0.5772\ldots$ is the Euler-Mascheroni constant. As in the case of $\nu < 1$, all moments of the distribution diverge in this case as well. See figure 2.3 for comparison with stochastic simulations.

The case of $\nu > 1$ requires a careful analysis of the function $\mathrm{J}(a, b; z)$ because the parameter a becomes negative in this regime. Since $\mathrm{J}(a, b; z)$ is the difference of beta and incomplete beta functions, the following identity holds

$$\mathrm{J}(a, b; z) = \frac{(a+b)}{a}\mathrm{J}(a+1, b; z) - \frac{z^a(1-z)^b}{a} \tag{2.24}$$

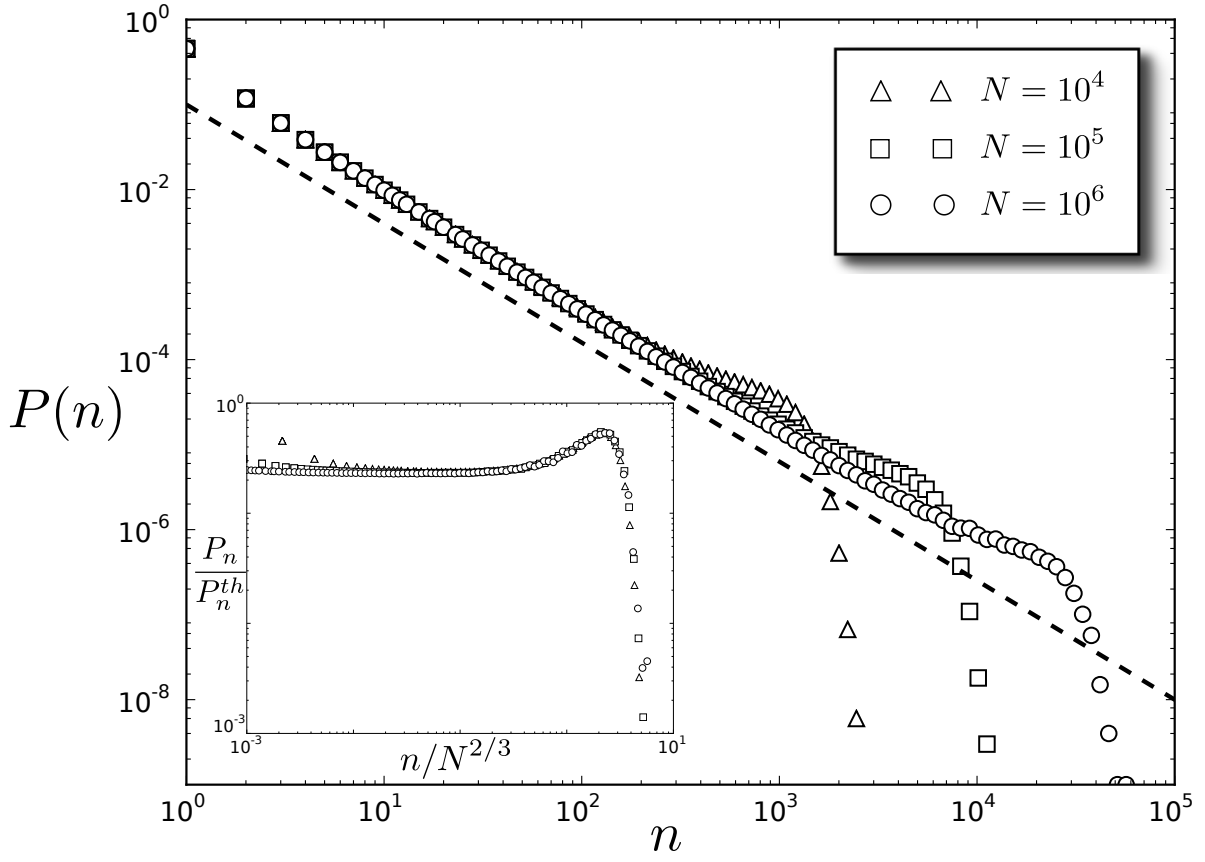


Figure 2.2: The probability of having an outbreak of size n empirically calculated from 10^9 stochastic realizations of the process (see appendix F) for different values of N and $\alpha = 1, \nu = 0.2$. The dashed line shows the slope of the analytical scaling predicted from theory (eq. 2.21), ignoring any constant prefactors. Inset shows the collapse of outbreak sizes when scaled by $N^{2/3}$. The bump near the exponential cutoff represents the probability mass associated with outbreaks that would have continued along the power-law in an infinite size system, but are clustered due to finite size effects.

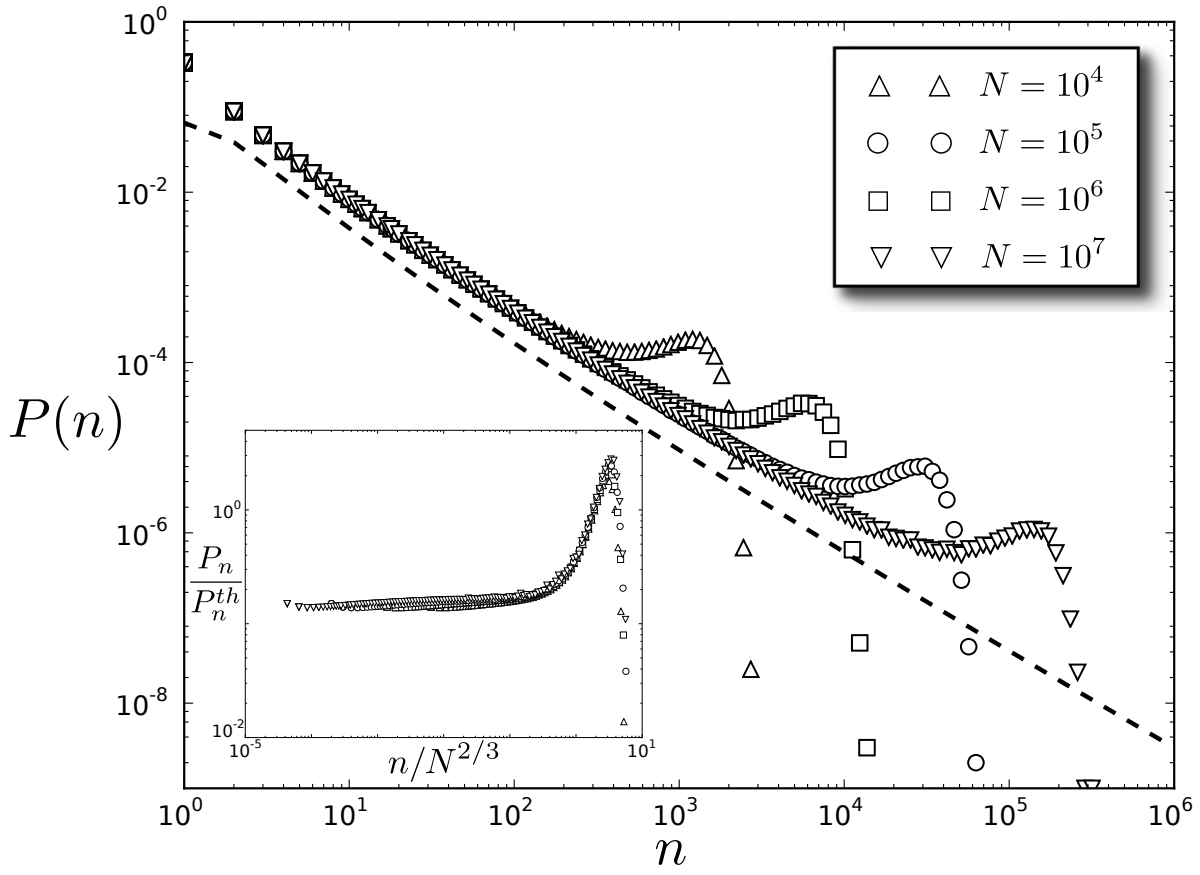


Figure 2.3: The probability of having an outbreak of size n for $\alpha = 1, \nu = 1$. Dashed line plotted at an offset represents the analytical scaling (eq. 2.23). Inset shows the collapse of outbreak sizes when scaled by $N^{2/3}$.

Consider the case where $\nu \in (1,2)$ which implies that $0 < a+1 < 1$. In this case as $z \to 0$,

$$\begin{aligned} \mathrm{J}(a+1,b;z) &\sim \mathrm{B}(a+1,b) \\ \mathrm{J}(a,b;z) &\sim \frac{(a+b)}{a}\mathrm{B}(a+1,b;z) - \frac{z^a(1-z)^b}{a} \end{aligned} \tag{2.25}$$

and the PGF $H(y)$ simplifies as follows

$$\begin{aligned} H(y) &\sim 1 + \frac{a}{\nu}\frac{\Lambda_1}{[1-(a+b)\mathrm{B}(a+1,b)z^{-a}(1-z)^{-b}]} \\ &= 1 + \frac{a\Lambda_1}{\nu}\left[1 + (a+b)\mathrm{B}(a+1,b)z^{-a}(1-z)^{-b} + \cdots\right] \end{aligned} \tag{2.26}$$

Substituting for z, we obtain a series expansion in fractional powers of $\sqrt{1-y}$. The leading term in the expansion is of the order $(1-y)^{(\nu-1)/2}$ which provides the following asymptotic form for $P(n)$.

$$P(n) \sim n^{-(\nu+1)/2} \tag{2.27}$$

Similarly, by binning ν in $\{(2,3),(3,4),\cdots\}$ and applying the property given by eq. 2.24 iteratively, we obtain the asymptotics as in eq. 2.27. The same is true when integral values are chosen for ν, in which case the PGF can be simplified further. For instance, substituting $\nu = 2$ in eq. 2.17a simplifies the PGF to

$$H(y) = \frac{1-\sqrt{1-y}}{2} \tag{2.28}$$

which is same as the PGF for the BD process with a prefactor of $1/2$. Asymptotic analysis reveals scaling exponent of $3/2$ consistent with eq. 2.27. The agreement of eq. 2.27 with stochastic simulations is demonstrated in figure 2.4.

Equation 2.27 suggests that the outbreak size distribution falls off more steeply with increasing ν. This seems counterintuitive at first because one would expect that, with increasing ν, there should be a greater probability of larger outbreaks, leading to more

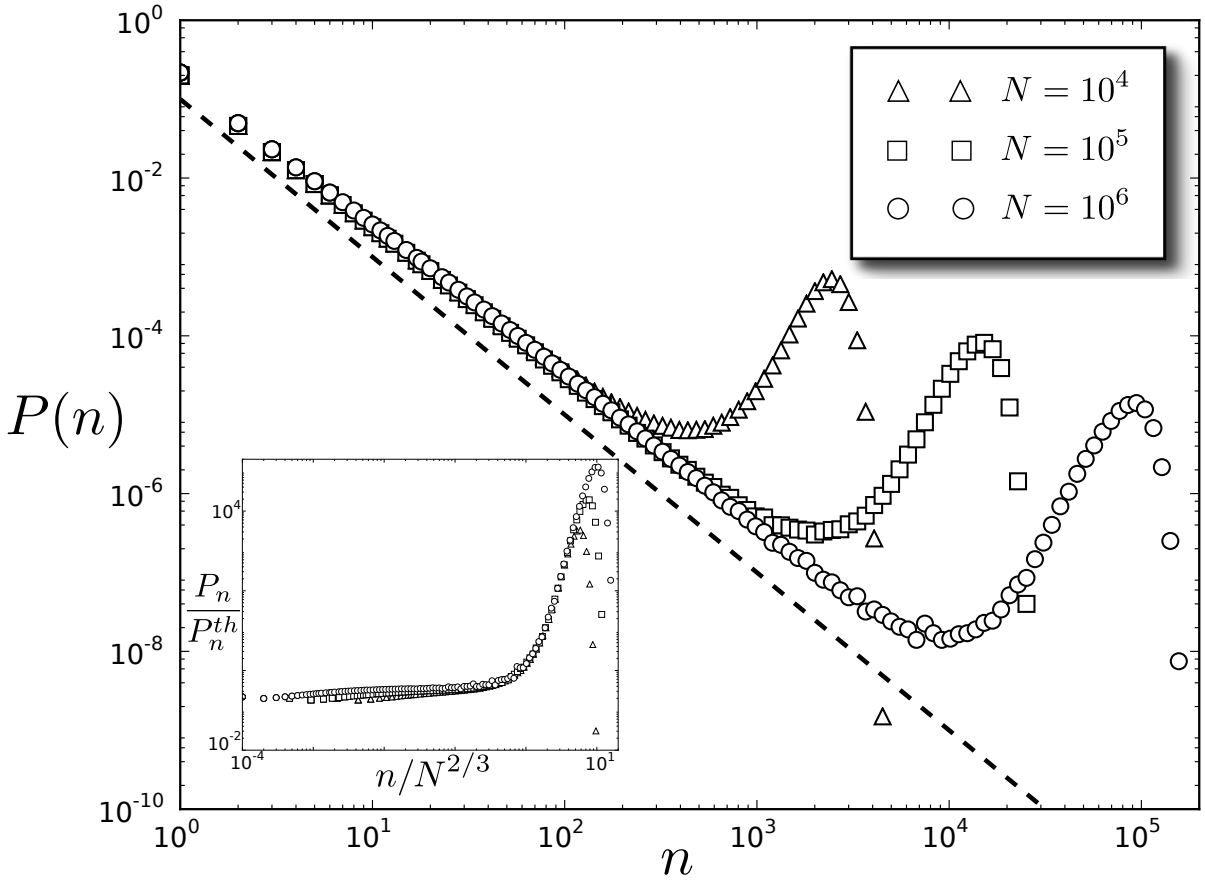


Figure 2.4: The probability of having an outbreak of size n for $\alpha = 1, \nu = 3$. Dashed line represents the analytical scaling (eq. 2.27) at an offset. The finite size bump is more pronounced because it has accumulated outbreaks of two distinct scales. The bump starts at the scale of $N^{2/3}$ consistent with what is observed for $\nu \leq 1$. However, the exponential cutoff which marks the end of the bump occurs at a different scale as evident from the lack of scaling collapse for large n in the inset. The presence of two scales in the bump is discussed in detail in section 2.3.1 and illustrated in figure 2.6.

slowly decaying distribution. The resolution of this puzzle can be found by looking at the total probability mass contained in the generating function. It can be verified that for $\nu \leq 1$, $H(y) \to 1$ as $y \to 1$, i.e., the distribution is proper. But for $\nu > 1$, the distribution becomes defective such that

$$\lim_{y \to 1} H(y) = \nu^{-1} \tag{2.29}$$

The remaining probability mass $1 - \nu^{-1}$ is associated with the infinite sized outbreak. This effect can be seen in stochastic simulations (figure 2.4) where the outbreaks not accounted by the power-law cluster in the bump of the distribution. More formally, we have the following

$$\lim_{n \to \infty} \mathbb{P}\big[R(\infty) > n\big] \sim \begin{cases} \mathcal{O}\left(n^{-(1-\nu)/2}\right) & \nu < 1, \\ \mathcal{O}\left((\log n)^{-1}\right) & \nu = 1, \\ 1 - \nu^{-1} + \mathcal{O}\left(n^{-(\nu-1)/2}\right) & \nu > 1. \end{cases} \tag{2.30}$$

Thus, for $\nu > 1$ the probability of having an outbreak size exceeding any arbitrary scale converges to a constant value of $1 - \nu^{-1}$ whereas the same probability diminishes with n for the case of $\nu \leq 1$. For $\nu > 1$, the distribution represented by the generating function $H(y)$ excludes the infinite sized outbreak. Thus, the statistics are conditional on a finite sized outbreak. The distribution falls more steeply with increasing ν because more and more outbreaks escape to infinity with probability $1 - \nu^{-1}$. Nevertheless, the statistics of the power-law regime are interesting to analyze even if they represent part of the distribution. For instance, the k^{th} moment of the distribution is finite only if $\nu > 2k + 1$ and diverges otherwise.

$$\langle n^k \rangle \sim \begin{cases} \dfrac{1}{\nu - (2k+1)} & \text{if } \nu > 2k + 1, \\ \infty & \text{otherwise.} \end{cases} \tag{2.31}$$

Finally, we summarize the asymptotic statistics calculated in the preceding section

$$P(n) \sim \begin{cases} n^{-(3-\nu)/2} & \nu < 1, \\ \dfrac{1}{n \log^2 4n} \left[\dfrac{1}{2} - \dfrac{\gamma}{\log 4n} + \mathcal{O}\left(\dfrac{1}{\log^2 4n} \right) \right] & \nu = 1, \\ n^{-(\nu+1)/2} & \nu > 1. \end{cases} \tag{2.32}$$

We can now put the results in some perspective. The external driving can be thought of as a 'coupling agent' that combines an increasing number of micro-outbreaks into a single outbreak as ν is increased. When $\alpha = 1$ and ν is above 1, the external driving binds an infinite number of micro-outbreaks into one contiguous outbreak with probability $1 - \nu^{-1}$. Qualitatively, the BDI process can also be interpreted as a two-state Markov chain that switches between the idle period and the busy period. In this interpretation the idle period is positive recurrent if $\alpha < 1$ (busy period ends with probability 1 and in finite time), null recurrent if $\alpha = 1, \nu \leq 1$ (busy period ends with probability 1 but the expected duration is ∞) and transient if $\alpha = 1, \nu > 1$ (busy period can persist indefinitely). The case of $\alpha > 1$ is trivial since a supercritical process can grow exponentially even without the external forcing. The idle period is thus transient in this case.

2.3 Finite population

2.3.1 Outbreak size

For a finite-sized system (eq. 2.2), we first establish the scaling of the 'maximal' outbreak size echoing the analysis in [15, 16]. Let there be a maximal size M, such that the outbreak

can not exceed this size due to depletion in the susceptible pool. For $\nu < 1$, the algebraic distribution in eq. 2.21 gives an estimate for the average outbreak size:

$$\langle n \rangle = \sum_{n \leq M} n \cdot P(n) \sim \sum_{n \leq M} n^{-(1-\nu)/2} \sim M^{(1+\nu)/2} \tag{2.33}$$

In a finite-sized system, the effective rate of infectious contact per infected host is reduced to $\alpha_\star = 1 - M/N$ due to depletion. From eq. 2.9, we obtain a second estimate for the scaling of the average outbreak size:

$$\langle n \rangle \sim (1 - \alpha_\star)^{-(1+\nu)} = (N/M)^{1+\nu}. \tag{2.34}$$

Equating the two estimates we obtain the following scaling laws for $\nu < 1$:

$$M \sim N^{2/3}, \quad \text{and} \quad \langle n \rangle \sim N^{(1+\nu)/3} \tag{2.35}$$

The scaling of M is verified in figure 2.2 (inset) and that of $\langle n \rangle$ in figure 2.7. For $\nu = 1$, we use the expression for $P(n)$ in eq. 2.23 and obtain M as the solution of the following implicit equation:

$$\left(\frac{N}{M}\right)^2 = \frac{M}{2\log^2 4M} + \mathcal{O}\left(\frac{M}{\log^3 4M}\right) \tag{2.36}$$

whose solution to a first order approximation leads to the following scaling laws

$$M \sim (N \log N)^{2/3} \quad \text{and} \quad \langle n \rangle \sim \left(\frac{N}{\log^2 N}\right)^{2/3} \tag{2.37}$$

However, numerical results obtained from stochastic simulations reveal slightly different scaling laws

$$M \sim (N^2 \log N)^{1/3}, \quad \text{and} \quad \langle n \rangle \sim \left(\frac{N}{\log N}\right)^{2/3} \tag{2.38}$$

that differ from theory by a factor of $(\log N)^{1/3}$ in M and $(\log N)^{-2/3}$ in $\langle n \rangle$. The empirical scaling law can be obtained if eq. 2.36 is replaced with the following

$$\left(\frac{N}{M}\right)^2 \sim \frac{M}{\log M} \tag{2.39}$$

Although the power-law part of the scaling – that is the term $N^{2/3}$ – is consistent between both the empirically observed (eq. 2.38) and the theoretically calculated (eq. 2.37) scaling, we are unable to resolve the logarithmic corrections and pose their solution as an open problem. The agreement of the scaling laws (eq. 2.38) with results from stochastic simulations is shown in figures 2.5 and 2.7. Henceforth, we shall refer only to the empirical scaling law for $\nu = 1$ where the logarithmic corrections are important.

The case of $\nu > 1$ requires careful consideration. The analysis on the infinite-sized system revealed that outbreaks occur according to a power law distribution (eq. 2.27) with probability ν^{-1} or are infinite in size with probability $1 - \nu^{-1}$. Henceforth, we shall label these as the 'power-law regime' and the 'infinite regime', respectively. The average outbreak size in the power-law regime diverges when $\nu < 3$ (see eq. 2.31). For finite systems, we expect that both the infinite regime and the power-law regime would admit two different scaling laws for average outbreak size and duration. The power-law regime admits a positive exponent for the scaling law only for $\nu \in (1, 3)$. Let $\langle n \rangle_\infty$ be the average outbreak size conditioned on the outbreak being in the infinite regime. From eq. 2.33, note that as $\nu \to 1$, the average outbreak size $\langle n \rangle$ approaches M in scale. For $\nu = 1$, we found empirically that $\langle n \rangle \sim M / \log M$ (see eq. 2.39). For $\nu > 1$, intuition suggests that in the infinite regime $\langle n \rangle_\infty \sim M$, i.e., all outbreaks will be clustered at one scale. Using eq. 2.34, we obtain the following scaling relationship

$$\langle n \rangle_\infty \sim M \sim N^{(\nu+1)/(\nu+2)} \tag{2.40}$$

The exponent of the scaling law in eq. 2.40 is an increasing function of ν that lies in the interval $(2/3, 1)$ for all $\nu > 1$. The lower bound of $2/3$ is consistent with the fact that the scaling law for $\nu < 1$ has $2/3$ as the upper bound (eq. 2.35) and that the same exponent shows up at $\nu = 1$ albeit with logarithmic factors (eq. 2.38). But the above scaling law will

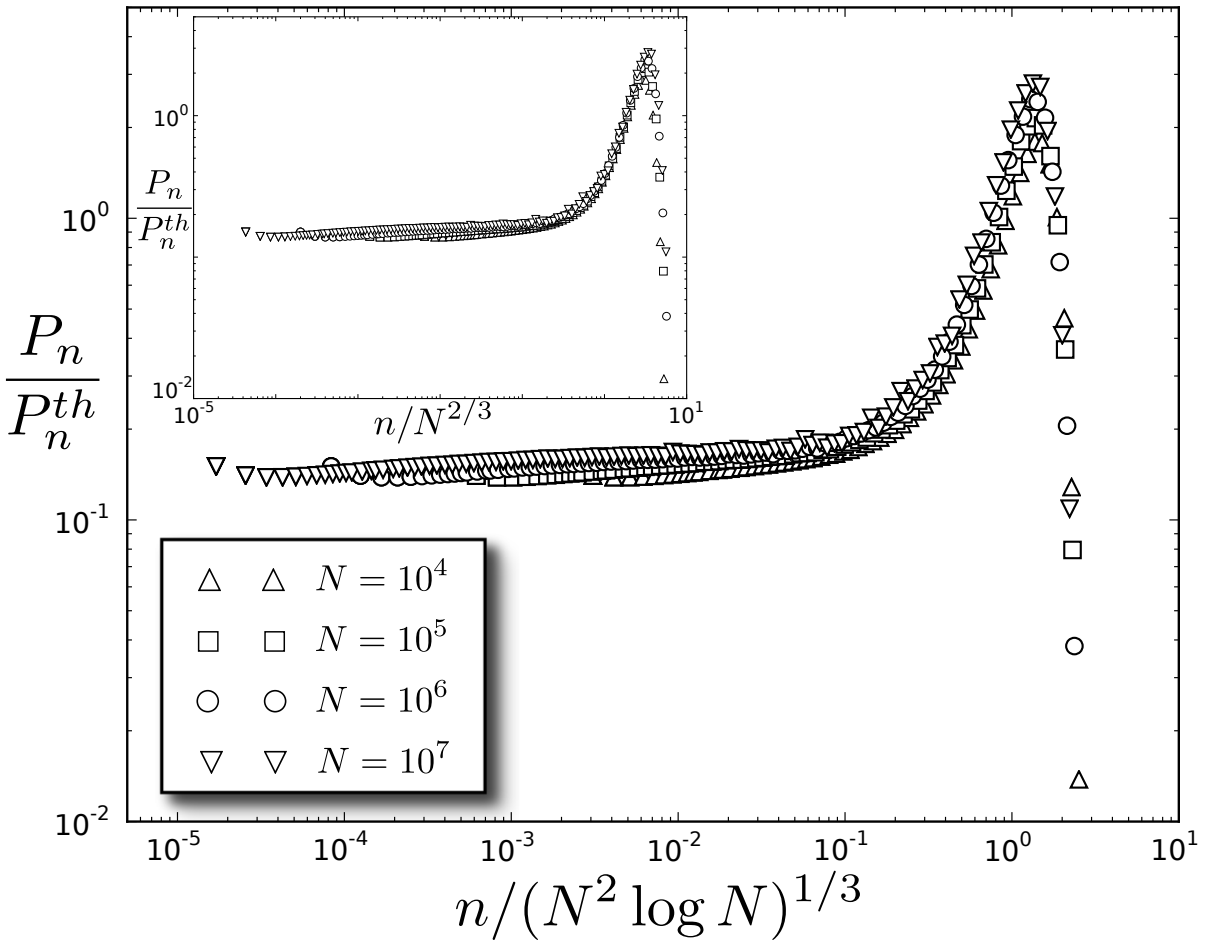


Figure 2.5: Scaling collapse at $M \sim (N^2 \log N)^{1/3}$ for $\nu = 1$. Y-axis is scaled by the theoretical scaling law of eq. 2.23. Note that the scaling collapse is distinct from the one shown in inset (done at the scale of $N^{2/3}$). While the power-laws collapse on top of each other at $N^{2/3}$, the exponential cutoffs collapse at $(N^2 \log N)^{1/3}$. This separation of scales is more pronounced for $\nu > 1$ (see figure 2.6).

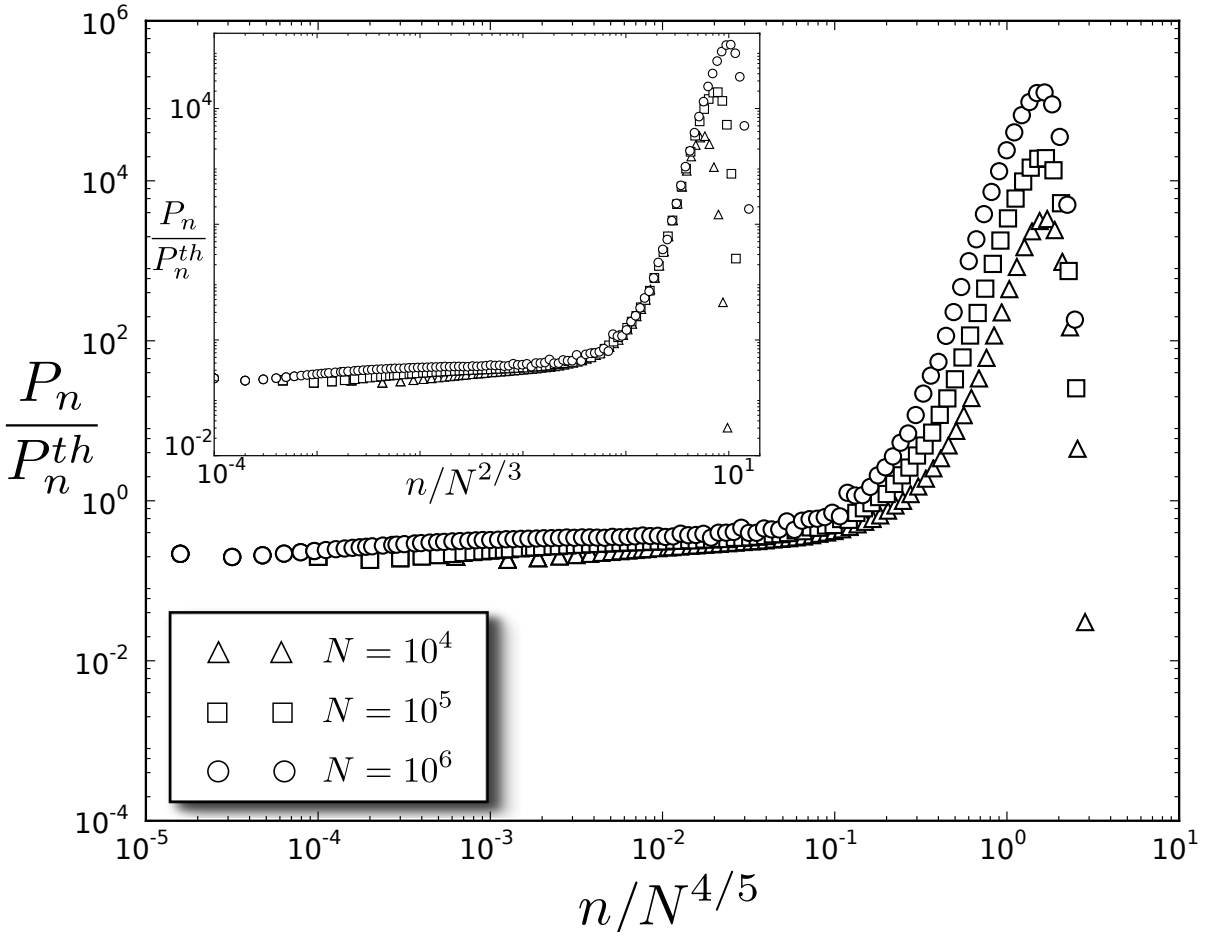


Figure 2.6: Scaling of the outbreak size distribution by $M \sim N^{(\nu+1)/(\nu+2)}$ for $\nu = 3$. The y-axis is scaled by the theoretical scaling law of eq. 2.27. Similar to figure 2.5, the scaling collapse is distinct from the one shown in inset (outbreaks scaled by $N^{2/3}$). The power-law regime exhibit a scaling collapse at $N^{2/3}$, while the exponential cutoffs collapse at $N^{4/5}$.

hold with probability $1 - \nu^{-1}$ that corresponds to the infinite regime. Let $\langle n \rangle_{pl}$ be the average outbreak size in the power-law regime. In a finite-sized system, there will be another scale L up to which the power-law regime holds, and any outbreak exceeding that enters the infinite regime. Using eq. 2.27, we can estimate the scaling in the power-law regime.

$$\langle n \rangle_{pl} \sim \sum_{n \leq L} n^{(1-\nu)/2} \sim \begin{cases} L^{(3-\nu)/2} & \nu \in (1, \infty) \backslash \{3\}, \\ \log L & \nu = 3. \end{cases} \tag{2.41}$$

L can be deduced by noting that in the limit of $\nu \to 1$, the scaling law (eq. 2.41) has to approach $N^{2/3}$ in order for the exponent to be consistent with the scaling laws for $\nu \leq 1$ (eq. 2.35 and 2.38). This is true only when L scales as the following

$$L \sim N^{2/3} \tag{2.42}$$

and thus we arrive at the following scaling laws,

$$\langle n \rangle_{pl} \sim \begin{cases} N^{1-\nu/3} & \nu \in (1, \infty) \backslash \{3\}, \\ \log N & \nu = 3 \end{cases} \tag{2.43}$$

The scale of $N^{2/3}$ as being the upper bound of all power-laws in $P(n)$ is confirmed in simulations (see inset in figures 2.2, 2.3 and 2.4). The derivation in eq. 2.43 assumes that L only depends on N and not ν. Intuitively, L is the scale at which the power-law regime is impacted by the finiteness of the system and thus should only depend on N. Other model parameters only determine how fast or slow the process approaches that scale. We now

summarize the finite-size scaling laws for the average outbreak size:

$$\langle n \rangle \sim \begin{cases} N^{(1+\nu)/3} & \nu < 1, \\ \left(\dfrac{N}{\log N}\right)^{2/3} & \nu = 1, \\ \left\{ \begin{array}{lll} N^{1-\nu/3} & \text{w.p.} & \nu^{-1} \\ N^{(\nu+1)/(\nu+2)} & \text{w.p.} & 1-\nu^{-1} \end{array} \right\} & \nu \in (1,\infty)\backslash\{3\}, \\ \left\{ \begin{array}{lll} \log N & \text{w.p.} & 1/3 \\ N^{4/5} & \text{w.p.} & 2/3 \end{array} \right\} & \nu = 3. \end{cases} \tag{2.44}$$

where w.p. is an abbreviation for 'with probability'. The agreement of these results with stochastic simulations is shown in figure 2.7. Similarly, the summary table for the maximal outbreak size is shown below.

$$M \sim \begin{cases} N^{2/3} & \nu < 1, \\ (N^2 \log N)^{1/3} & \nu = 1, \\ N^{(\nu+1)/(\nu+2)} & \nu > 1. \end{cases} \tag{2.45}$$

The critical point of $\nu = 1$ separates the scaling behavior of M into one as being a power law with fixed exponent of $2/3$ and the other as a power law with continuously varying exponent.

For $\nu > 1$, the scaling exponent of the average outbreak size bifurcates at the value of $2/3$; the two different exponents move in opposite directions with increasing ν (compare $1 - \nu/3$ with $(\nu+1)/(\nu+2)$ both of which start off from the value $2/3$ as $\nu \to 1+$). A crucial insight from these results is that the average outbreak size scales as N^{ξ} where $\xi \in (0,1)\backslash\{2/3\}$ (at $\xi = 2/3$, logarithmic corrections are present). The scaling law is always

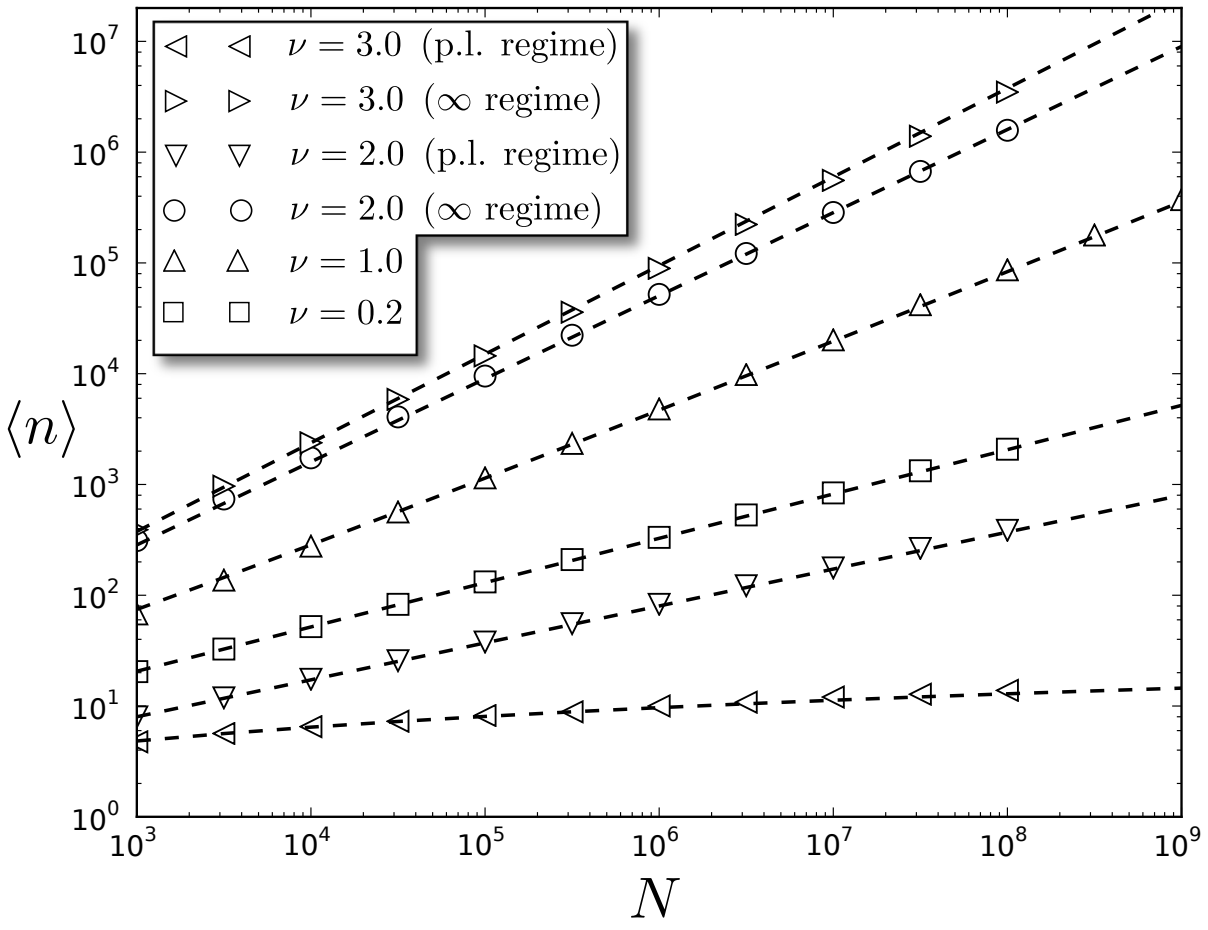


Figure 2.7: Finite size scaling for average outbreak size ($\nu = \{0.2, 1, 2, 3\}$). The statistics of the power law regime and the infinite regime were calculated separately using $N^{2/3}$ as the separation boundary. Note that the scaling for $\nu = 3$ is purely logarithmic in the power-law regime. Dashed lines represent the scaling laws predicted from theory (eq. 2.44).

sublinear as long as $\alpha = 1$, i.e., there cannot occur an outbreak that scales as $\mathcal{O}(N)$ no matter how strongly the system is driven. Only for $\alpha > 1$, would there be an $\mathcal{O}(N)$ outbreak with ν having no qualitative bearing on the statistics. This is because the multiplicative nature of the supercritical BD process always dominates the constant rate of growth from external driving.

Using the above results, we can calculate the scaling window for the scaling laws, i.e., the distance from the threshold boundary within which the scaling laws are applicable [16]. The scaling window is a characteristic of the finite system size and shrinks to 0 in the limit of $N \to \infty$. For finite N, the system need not be right at the critical threshold $\alpha = 1$ for the scaling laws to be valid. Using eq. 2.34 and eq. 2.44 we obtain,

$$|\alpha - 1| \sim \begin{cases} N^{-1/3} & \nu < 1, \\ \left(\dfrac{N}{\log N}\right)^{-1/3} & \nu = 1, \\ \begin{cases} N^{-1/3} & \text{power law regime} \\ N^{-1/(\nu+2)} & \text{infinite regime} \end{cases} & \nu > 1. \end{cases} \tag{2.46}$$

For a fixed N, the infinite regime has the largest window that grows with ν.

2.3.2 Outbreak duration

With the effective transmission rate $\alpha_\star = 1 - M/N$ below 1, the scaling behavior for outbreak durations can be obtained by using eq. 2.10:

$$\begin{aligned}\langle t \rangle &= \frac{(1-\alpha_\star)^{-\nu} - 1}{\nu \alpha_\star} \\ &\sim \begin{cases} \log(N/M) & \nu = 0, \\ (N/M)^\nu & \nu > 0. \end{cases}\end{aligned} \tag{2.47}$$

For $\nu \leq 1$, we arrive at the following using eq. 2.45

$$\langle t \rangle \sim \begin{cases} \log N & \nu = 0, \\ N^{\nu/3} & 0 < \nu < 1, \\ \left(\dfrac{N}{\log N}\right)^{1/3} & \nu = 1. \end{cases} \tag{2.48}$$

For $\nu > 1$, we have the bifurcation of behavior into the power law regime and the infinite regime. Since we already know the scale of M in the infinite regime (from eq. 2.45), we obtain

$$\langle t \rangle_\infty \sim N^{\nu/(\nu+2)}, \qquad \nu > 1. \tag{2.49}$$

In the power law regime, we resort to the survival function for calculating the scaling for the average outbreak duration (see Appendix C.2) and obtain the following

$$\langle t \rangle_{pl} \sim \begin{cases} T_c^{2-\nu} & \nu \in (1, \infty) \backslash \{2\} \\ \log T_c & \nu = 2 \end{cases} \tag{2.50}$$

where T_c is the cutoff timescale for the power law regime. For $\nu > 1$, we know that the cutoff length scale for the power law is $L \sim N^{2/3}$. We now estimate the relationship

between L and T_c. From eq. 2.11b, the mean number of infectious hosts increases linearly with time. Thus, the outbreak size grows quadratically with time,

$$\frac{dR}{dt} \sim t, \quad R \sim t^2 \tag{2.51}$$

and this gives the relationship between T_c and L as

$$T_c \sim \sqrt{L} \tag{2.52}$$

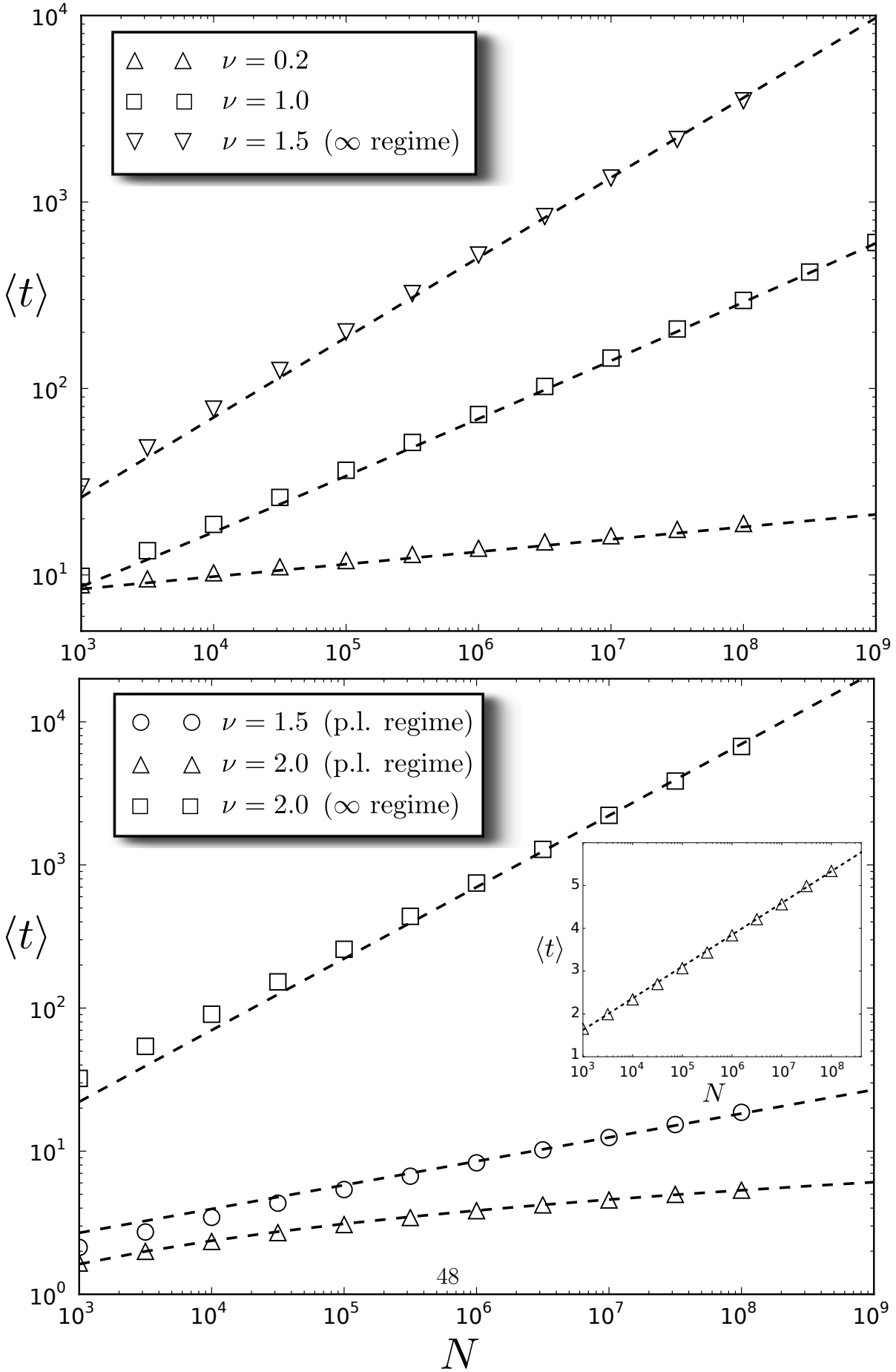
The same scaling relationship was noted in [16] for the simple SIR. Using eq. 2.50, 2.52 and 2.42, we obtain

$$T_c \sim N^{1/3}, \quad \langle t \rangle_{pl} \sim \begin{cases} N^{(2-\nu)/3} & \nu \in (1,\infty)\backslash\{2\}, \\ \log N & \nu = 2. \end{cases} \tag{2.53}$$

The summary of the scaling laws for the average outbreak duration is given below and the agreement with stochastic simulations is shown in figure 2.8.

$$\langle t \rangle \sim \begin{cases} \log N & \nu = 0, \\ N^{\nu/3} & \nu < 1, \\ \left(\dfrac{N}{\log N}\right)^{1/3} & \nu = 1, \\ \begin{cases} N^{(2-\nu)/3} & \text{w.p.} \quad \nu^{-1} \\ N^{\nu/(\nu+2)} & \text{w.p.} \quad 1-\nu^{-1} \end{cases} & \nu \in (1,\infty)\backslash\{2\}, \\ \begin{cases} \log N & \text{w.p.} \quad 1/2 \\ N^{1/2} & \text{w.p.} \quad 1/2 \end{cases} & \nu = 2. \end{cases} \tag{2.54}$$

On comparing the scaling laws for the average outbreak size (eq. 2.44) and duration (eq.


$\nu = 0.2$
$\nu = 1.0$
$\nu = 1.5$ (∞ regime)
$\langle t \rangle$
$\nu = 1.5$ (p.l. regime)
$\nu = 2.0$ (p.l. regime)
$\nu = 2.0$ (∞ regime)
N

2.54), we note that in all power law regimes ($\nu \notin \{0, 1, 2, 3\}$), the following relationship holds

$$\frac{\langle n \rangle}{\langle t \rangle} \sim N^{1/3} \tag{2.55}$$

For $\nu \in \{0, 1, 2, 3\}$, the relationship is not too far off either with the presence of logarithmic factors,

$$\frac{\langle n \rangle}{\langle t \rangle} \sim \begin{cases} \dfrac{N^{1/3}}{\log N} & \nu = 0, \\ \left(\dfrac{N}{\log N}\right)^{1/3} & \nu = 1, \\ \dfrac{N^{1/3}}{\log N} & \nu = 2, \\ N^{1/3} \log N & \nu = 3. \end{cases} \tag{2.56}$$

In the infinite regime however (only for $\nu > 1$),

$$\frac{\langle n \rangle_\infty}{\langle t \rangle_\infty} \sim N^{1/(\nu+2)} \tag{2.57}$$

the two scales converge for fixed N and increasing ν. This result shows that there is a universality in the power-law characteristics regardless of whether ν is below or above the critical value of 1, and it ties with the universality of $N^{2/3}$ as the characteristic scale to which all power laws in $P(n)$ extend. The presence of an 'infinite regime' does not preclude this universality.

2.3.3 Convergence near critical points

The finite-size scaling laws come with a caveat: that the system size should be large enough or the parameter ν should be far away from critical points to avoid any interference from the logarithmic factors (see eq. 2.44 and 2.54). For instance, if $\nu = 1 \pm \epsilon$, the logarithmic

factor present in the scaling laws for $\nu = 1$ interferes with the scaling laws for $\nu < 1$ and $\nu > 1$ if $\epsilon \ll 1$. Since all scaling laws in the infinite regime have a monotonically increasing exponent, they would not be subjected to any interference near the critical points. Similar to eq. 2.46, we can calculate heuristically, the 'interference window' for ν within which scaling power laws will be muddied via interference from logarithmic factors. Near $\nu = 1$, interference would occur if the two estimates of $\langle n \rangle$ at and above $\nu = 1$ are similar in scale, i.e.,

$$\left(\frac{N}{\log N}\right)^{2/3} \sim N^{1-\nu/3} \tag{2.58}$$

which gives the interference window as

$$|\nu - 1| \sim \frac{\log(\log N)}{\log N} \tag{2.59}$$

The window is a slowly decreasing function of N. The same functional form is obtained if we compare the scale of $\langle n \rangle$ at and below $\nu = 1$, as well as near all other critical points for both $\langle n \rangle$ and $\langle t \rangle$, i.e.,

$$|\nu - \nu_c| \sim \frac{\log(\log N)}{\log N} \tag{2.60}$$

Thus, for moderate values of N, the scaling laws for the power-law regime are likely to suffer from interference from logarithmic factors unless ν is far away from its critical values.

2.4 Discussion

In this work, we have solved for the statistical properties of the externally forced SIR model through rigorous analysis of the relevant stochastic process. By invoking the analogy between the BDI process and the $M/G/\infty$ queue, we were able to leverage existing results in building the theory for the process. The external driving acts as a binding agent for micro-outbreaks and is especially significant when $\alpha = 1$. In this case $\nu = 1$ emerges as a

second critical point in the process separating a state of recurring outbreaks from one with a single perpetual outbreak. Although power-law characteristics at the critical point were expected, the tunability of the power law by the external forcing with a precise functional form is a non-trivial result that was revealed through calculations. The finite-size scaling laws exhibit a continuum of scaling exponents governed by the driving rate that has some important implications for understanding reservoir-driven epidemics. This work also elucidates the universality of the scale of the maximal outbreak size and of the ratio of average outbreak size and duration when the distribution of sizes $P(n)$ follows a power law.

Our results provide a framework for interpreting time series data from reservoir-driven outbreaks where the timescale of primary infections (direct reservoir transmission) and secondary infections (transmission among hosts) are comparable, and where it is not feasible to conduct field studies necessary to distinguish among them. In cases of sufficiently weak reservoir forcing, individual chains of secondary transmission can be explained by simple SIR statistics. But we demonstrate here that if the system is near the critical threshold which is typical of emerging infectious diseases, the statistics of the process depend strongly on the reservoir forcing. Similarly, if fine scale data were available that allowed one to resolve each micro-outbreak separately, then the simple SIR process is sufficient to describe the data. But typically, such fine scale data are difficult to collect on the timescales of outbreaks, and practitioners often have to contend with coarse scale data on composite outbreaks, which is precisely where our theory and results serve a strong purpose.

CHAPTER 3

Chapter 3: Using multitype branching processes to quantify statistics of disease outbreaks in zoonotic epidemics

ABSTRACT [1]

Branching processes have served as a model for chemical reactions, biological growth processes and contagion (of disease, information or fads). Through this connection, these seemingly different physical processes share some common universalities that can be elucidated by analyzing the underlying branching process. In this work we focus on coupled branching processes as a model of infectious diseases spreading from one population to another. An exceedingly important example of such coupled outbreaks are zoonotic infections that spill over from animal populations to humans. We derive several statistical quantities characterizing the first spillover event from animals to humans, including the probability of spillover, the first passage time distribution for human infection, and disease prevalence in the animal population at spillover. Large stochastic fluctuations in those quantities can make inference of the state of the system at the time of spillover difficult. Focusing on outbreaks in the human population, we then characterize the critical threshold

[1] This chapter and the material in Appendix D are published in *Physical Review E* with co-authors David J. Schneider and Christopher R. Myers [29].

for a large outbreak, the distribution of outbreak sizes, and associated scaling laws. These all show a strong dependence on the basic reproduction number in the animal population, and indicate the existence of a novel multicritical point with altered scaling behavior. The coupling of animal and human infection dynamics has crucial implications, most importantly allowing for the possibility of large human outbreaks even when human-to-human transmission is subcritical.

3.1 Introduction

Coupled reaction processes taking place across spatial domains or populations with complex structure can exhibit rich dynamics and phase transitions. Reaction-diffusion models have been used to study wave propagation and pattern formation [38], and epidemic models have been useful in understanding the spread of infectious diseases, rumors, computer viruses and fads in populations [24]. In the context of stochastic formulations, branching processes often form the basis of such models. In describing epidemics, the Susceptible-Infected-Recovered (SIR) model has been extensively studied in fully-mixed populations [26], and more recently on complex networks [39, 40]. The SIR model of disease dynamics is an important cornerstone of mathematical epidemiology where systematic analysis has been possible through the use of branching processes [12, 26]. In the limit of large population, the SIR process converges to a linear birth-death process — a special case of continuous time branching processes [11]. The linearized process is amenable to analytical treatment that yields some important results about the original nonlinear process, such as the existence of a second order phase transition at a critical threshold. Below the threshold, all outbreaks are small (with size $o(N)$ as population size $N \to \infty$) whereas above the threshold some outbreaks can be large (with size $\mathcal{O}(N)$ as $N \to \infty$). At the critical point, the distribution

of outbreak sizes shows a power-law scaling of $P(n) \sim n^{-3/2}$ (where $P(n)$ is the probability of an outbreak of size n) with the average outbreak size scaling as $N^{1/3}$ [16].

Whereas the structure and statistics of epidemics in single population are extremely well characterized, the structure of coupled epidemics in metapopulations have received much less attention. An exceedingly important example of such coupled outbreaks are zoonotic infections that spill over from animal populations to humans representing a major challenge in public health [27, 41, 42]. A zoonotic disease system typically involves one or more animal species with humans as the end hosts where cross-species transmission (*spillover*) is facilitated by direct or vector-mediated interactions between animals and humans. Although recent work has sought to characterize and classify the salient features of zoonoses [27, 43, 44], such classification schemes are largely descriptive, and the basic phenomenology of cross-species infection has not been addressed in sufficient mathematical detail. Models that explicitly incorporate spillover dynamics are exceedingly rare, despite the fact that such events are the defining characteristic of zoonotic infection [27]. Among models that do exist, stochastic treatments of spillover dynamics are much less common than deterministic models, a fact echoed in a recent survey [45].

In a zoonotic system, it is important to investigate how the nature of the critical threshold and the statistics of human outbreaks, which are well known for the simple SIR model, change with the addition of spillover dynamics. We address this important question in this paper using a simple two-species model of zoonoses (see figure 3.1). The model at its core represents two SIR processes that are coupled through cross-species transmission. The two host populations, animal and human, are fully mixed within their respective species, with a partial overlap between the species. The partial overlap or the 'mixing fraction', ν, represents the fraction of human hosts that are fully mixed with the animal hosts, and are thus at risk for direct infection from animals. The three types of possible infection

transmission reactions are animal-to-animal, animal-to-human, and human-to-human. This model describes zoonoses where the infection prevalence in the animal population changes rapidly on timescales of interest. This might occur with the introduction of a disease into a new amplifier animal host population [46, 47] or with the evolutionary emergence of a new, more virulent strain of an existing animal pathogen [48, 49]. Due to the dynamical nature of spillover driven by the animal SIR process, the degree of animal-to-animal transmission becomes an important determinant of human outbreaks alongside the degree of human-to-human transmission.

While the statistics of human outbreaks are crucial for developing a systematic understanding, there are additional questions that one may ask specific to a cross-species disease system. These include the probability of spillover, the distribution of time to spillover (from the point when the infection starts in the animal population) and the distribution of infected animals at the moment of spillover. These questions are relevant from the perspective of parameter estimation and control interventions, as we address in this paper.

An outline of our methodology and results is as follows. The coupled SIR process (figure 3.1) converges to a multi-type linear birth-death process in the limit of large system size for fixed ratio of animal and human population size. The limiting linear process allows us to use the same techniques that have been employed to analyze the simple SIR in the limit of large system size. The results presented here are divided into two main parts. First, we calculate the joint distribution of a subset of state variables as a function of time using probability generating functions (PGF). We use the PGF to calculate the probability of spillover, the distribution of first passage times, and the distribution of infected animals at the time of spillover. Second, we calculate the PGF for the distribution of outbreak sizes in humans, which we then use to calculate the critical threshold, the scaling laws for the

distribution function at the critical point as well as finite size scaling for the average outbreak size. Lastly, we calculate the probability of a large human outbreak.

The basic reproduction number $\mathcal{R}_0$ for an epidemic is defined as the average number of new infections produced by an infectious host in a fully susceptible population. In our analysis we find that the critical threshold is a function of the basic reproduction numbers in both animals and humans, $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{hh}$, and identify a parameter regime where large outbreaks are possible in human populations – *sustained by repeated introductions from the animal population* – even if human-to-human transmission is subcritical (i.e., when $\mathcal{R}_0^{hh} < 1$). Information only about infection in the human population is insufficient to distinguish such a scenario from one involving a single primary introduction followed by extensive human-to-human transmission (see fig. 3.1 (b)). Our systematic characterization of the spectrum of possible behaviors helps to augment and clarify the previously proposed classification frameworks [27, 43, 44]. We see our work as a stepping stone toward more complex and realistic models that might help address the spatial and ecological aspects of zoonotic emergence, the evolution of virulence, and public health interventions in the form of dynamic control strategies.

3.2 The coupled SIR process

The three–type metapopulation model considered here consists of animals, type 1 humans and type 2 humans. Figure 3.1 shows a schematic of the model and the reaction equations

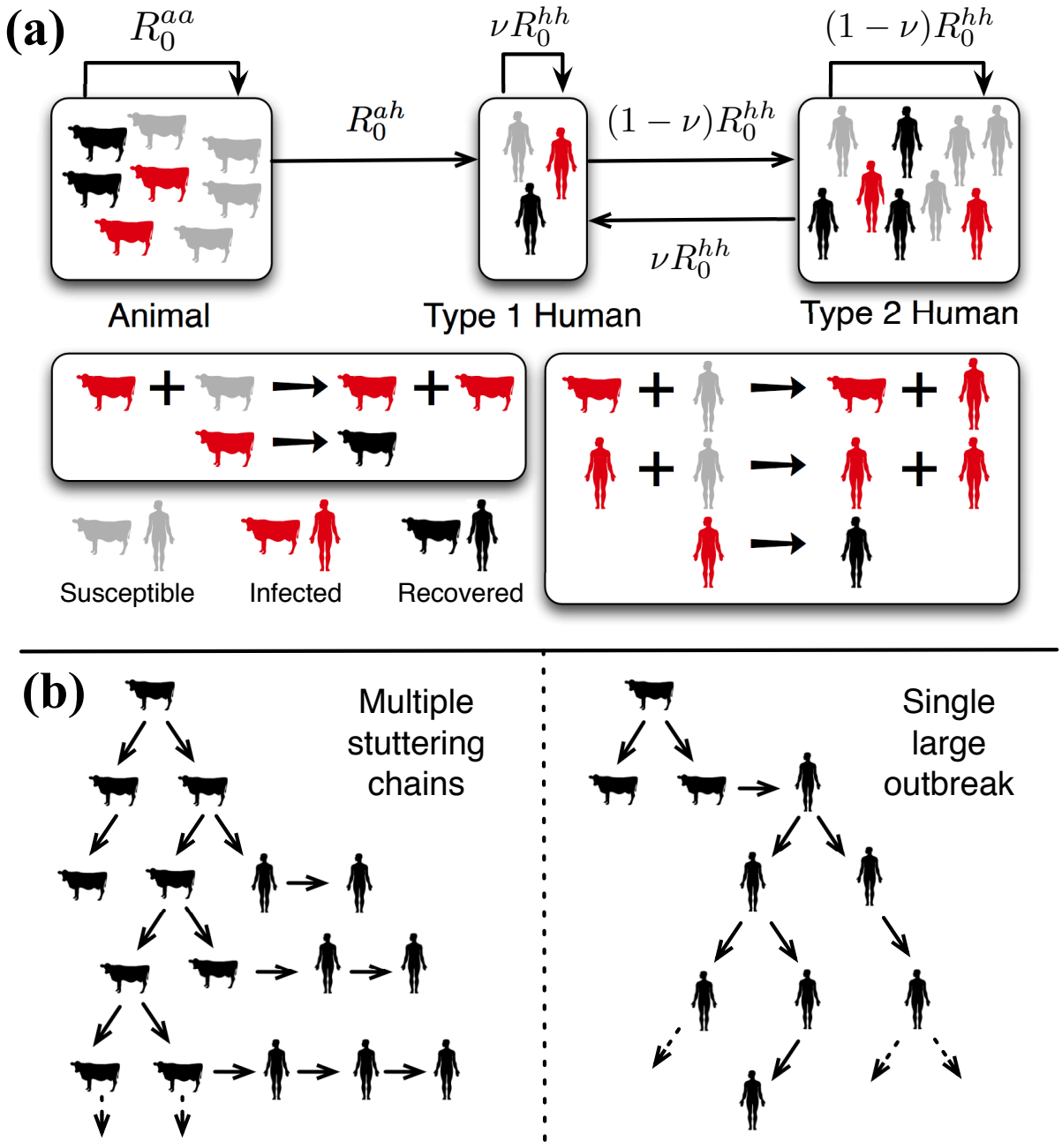


Figure 3.1: (Color online) **(a)** Schematic of our zoonoses model. The labeled arrows denote $\mathcal{R}_0$'s for inter- and intra-population transmission. **(b)** Schematic depicting two possible mechanisms for zoonotic outbreaks in human populations: (Left) Infection spreads efficiently in the animal population but inefficiently in humans, with each introduction into humans leading to a stuttering chain that goes extinct. (Right) An initial spillover leads to a large outbreak sustained by human-to-human transmission.

are as follows:

$$
\begin{aligned}
(S_a, I_a, \ldots) &\xrightarrow{\beta_{aa} S_a I_a / N_a} (S_a - 1, I_a + 1, \ldots) \\
(\ldots, I_a, R_a, \ldots) &\xrightarrow{\gamma_a I_a} (\ldots, I_a - 1, R_a + 1, \ldots) \\
(\ldots, S_{h1}, I_{h1p}, \ldots) &\xrightarrow{\beta_{ah} S_{h1} I_a / N_a} (\ldots, S_{h1} - 1, I_{h1p} + 1, \ldots) \\
(\ldots, S_{h1}, I_{h1s}, \ldots) &\xrightarrow{\beta_{hh} S_{h1} I_h / N_h} (\ldots, S_{h1} - 1, I_{h1s} + 1, \ldots) \\
(\ldots, I_{h1p}, R_{h1p}, \ldots) &\xrightarrow{\gamma_h I_{h1p}} (\ldots, I_{h1p} - 1, R_{h1p} + 1, \ldots) \\
(\ldots, I_{h1s}, R_{h1s}, \ldots) &\xrightarrow{\gamma_h I_{h1s}} (\ldots, I_{h1s} - 1, R_{h1s} + 1, \ldots) \\
(\ldots, S_{h2}, I_{h2}, \ldots) &\xrightarrow{\beta_{hh} S_{h2} I_h / N_h} (\ldots, S_{h2} - 1, I_{h2} + 1, \ldots) \\
(\ldots, I_{h2}, R_{h2}) &\xrightarrow{\gamma_h I_{h2}} (\ldots, I_{h2} - 1, R_{h2} + 1)
\end{aligned}
\tag{3.1}
$$

The full state vector is:

$$(S_a, I_a, R_a, S_{h1}, I_{h1p}, I_{h1s}, R_{h1p}, R_{h1s}, S_{h2}, I_{h2}, R_{h2})$$

but for brevity, only those state variables that change are shown in each reaction above. There are N_a animal hosts and N_h human hosts in the system. A fraction ν of all human hosts are of type 1 that can receive both primary transmission from animals and secondary transmission from other humans. Type 2 human hosts, which are a fraction $1 - \nu$ of the total human population, can only receive a secondary transmission. Type 1 humans are fully mixed with both animals and type 2 humans. $(S_\star, I_\star, R_\star)$ denote the population in the susceptible, infectious and recovered compartments of each group such that $S_\star + I_\star + R_\star = N_\star$. The subscripts (or superscripts) aa, ah and hh stand for the animal–animal, animal–human and human–human transmission reactions, respectively. β_{aa}, β_{ah} and β_{hh} are the rates of infectious contact per infectious host, and γ_a and γ_h represent the rates of recovery. The subscripts 'p' and 's' (such as in I_{h1p} and I_{h1s}) distinguish between primary and secondary infections, respectively. The total number of infected human hosts is given by $I_h = I_{h1p} + I_{h1s} + I_{h2}$.

The basic reproduction numbers associated with the aa, ah and hh transmissions are:

$$\mathcal{R}_0^{aa} = \frac{\beta_{aa}}{\gamma_a}, \quad \mathcal{R}_0^{hh} = \frac{\beta_{hh}}{\gamma_h}, \quad \mathcal{R}_0^{ah} = \frac{\nu\beta_{ah}}{\rho\gamma_a} \equiv \frac{\hat{\beta}_{ah}}{\gamma_a} \tag{3.2}$$

$\mathcal{R}_0^{ij}$ represents the average number of new infections produced by a single infected host of type i in a fully susceptible population of type j. The animal epidemic follows the simple SIR process with $\mathcal{R}_0^{aa} = 1$ as the critical threshold. For $\mathcal{R}_0^{aa} < 1$, all animal outbreaks will be small. For $\mathcal{R}_0^{aa} > 1$, the probability of having a large outbreak (of size $\mathcal{O}(N_a)$) is $1 - 1/\mathcal{R}_0^{aa}$.

3.3 The multitype linear birth-death process

We investigate the model in the limit of $N_a, N_h \to \infty, N_a/N_h \to \rho$. In this limit, the epidemic does not saturate and the depletion of susceptible pool in finite time can be ignored. The original nonlinear process reduces to a multitype linear birth-death process [10] with the following rate equations:

$$\begin{aligned}
&(I_a, \ldots) && \xrightarrow{\beta_{aa} I_a} (I_a{+}1, \ldots) \\
&(\ldots, I_a, R_a, \ldots) && \xrightarrow{\gamma_a I_a} (\ldots, I_a{-}1, R_a{+}1, \ldots) \\
&(\ldots, I_{h1p}, \ldots) && \xrightarrow{\nu\beta_{ah} I_a/\rho} (\ldots, I_{h1p}{+}1, \ldots) \\
&(\ldots, I_{h1s}, \ldots) && \xrightarrow{\nu\beta_{hh} I_h} (\ldots, I_{h1s}{+}1, \ldots) \\
&(\ldots, I_{h1p}, R_{h1p}, \ldots) && \xrightarrow{\gamma_h I_{h1p}} (\ldots, I_{h1p}{-}1, R_{h1p}{+}1, \ldots) \\
&(\ldots, I_{h1s}, R_{h1s}, \ldots) && \xrightarrow{\gamma_h I_{h1s}} (\ldots, I_{h1s}{-}1, R_{h1s}{+}1, \ldots) \\
&(\ldots, I_{h2}, \ldots) && \xrightarrow{(1-\nu)\beta_{hh} I_h} (\ldots, I_{h2}{+}1, \ldots) \\
&(\ldots, I_{h2}, R_{h2}) && \xrightarrow{\gamma_h I_{h2}} (\ldots, I_{h2}{-}1, R_{h2}{+}1)
\end{aligned} \tag{3.3}$$

The linear process can be represented by a smaller set of equations, if we ignore the type labeling. Before doing so, we introduce a new set of state variables for the human hosts that would be convenient for forthcoming analysis. Let $Z_\star(t) = I_\star(t) + R_\star(t)$ where $\star$ stands for particular subscripts used in what follows. $Z_{h,p}(t)$ denotes the number of primary human infections and $Z_{h,s}(t)$ denotes the number of secondary human infections irrespective of the human host type. The total number of infected human hosts is then $Z_h(t) = Z_{h,p}(t) + Z_{h,s}(t)$, regardless of type. The reduced set of equations describing the multitype linear birth-death process are:

$$\begin{aligned}
&(I_a, R_a, I_h, Z_{h,p}, Z_{h,s}) \xrightarrow{\beta_{aa} I_a} (I_a{+}1, R_a, I_h, Z_{h,p}, Z_{h,s}) \\
&(I_a, R_a, I_h, Z_{h,p}, Z_{h,s}) \xrightarrow{\gamma_a I_a} (I_a{-}1, R_a{+}1, I_h, Z_{h,p}, Z_{h,s}) \\
&(I_a, R_a, I_h, Z_{h,p}, Z_{h,s}) \xrightarrow{\hat{\beta}_{ah} I_a} (I_a, R_a, I_h{+}1, Z_{h,p}{+}1, Z_{h,s}) \\
&(I_a, R_a, I_h, Z_{h,p}, Z_{h,s}) \xrightarrow{\beta_{hh} I_h} (I_a, R_a, I_h{+}1, Z_{h,p}, Z_{h,s}{+}1) \\
&(I_a, R_a, I_h, Z_{h,p}, Z_{h,s}) \xrightarrow{\gamma_h I_h} (I_a, R_a, I_h - 1, Z_{h,p}, Z_{h,s})
\end{aligned} \tag{3.4}$$

where $\hat{\beta}_{ah} \equiv \nu\beta_{ah}/\rho$. The description of the process dynamics can be expressed in the form of probability $P_{i,j,k,l,m}(t)$ of the state variables $(I_a, R_a, I_h, Z_{h,p}, Z_{h,s})$ being in the state (i, j, k, l, m) at time t given that the process starts with a single infectious animal host. Similarly, let $Q_{n,p}(t)$ be the probability of the state variables $(I_h, Z_{h,s})$ being in the state (n, p) at time t starting from a single infectious human host at time 0. The set of all possible transitions that involve a single infectious animal host are: production of another infectious animal host, moving into recovery or production of an infectious human host. Let the probability of being in the state (i, j, k, l, m) at time t from these one step transitions be

$$P^{a\to aa}_{i,j,k,l,m}(t),\ P^{a\to\emptyset}_{i,j,k,l,m}(t) \text{ and } P^{a\to ah}_{i,j,k,l,m}(t)$$

respectively. Similarly, for the transitions that begin with a single infectious human host, let $Q^{h\to hh}_{n,p}$ and $Q^{h\to\emptyset}_{n,p}$ be the probability of reaching the (n, p) state based on infection or

recovery taking place at the first step. The probability $P^{a\to\emptyset}_{i,j,k,l,m}(t)$ is 1 when $i=k=0$ and 0 otherwise. Similarly, $Q^{h\to\emptyset}_{n,p}$ is 1 when $n=0$ and 0 otherwise. These probabilities would satisfy the Kolmogorov backward equation [11]:

$$\begin{aligned}\frac{dP_{i,j,k,l,m}}{dt} &= \beta_{aa}P^{a\to aa}_{i,j,k,l,m} + \gamma_a\delta_{i,0}\delta_{k,0} + \beta_{ah}P^{a\to ah}_{i,j,k,l,m} \\ &\quad - (\beta_{aa}+\gamma_a+\beta_{ah})P_{i,j,k,l,m} \\ \frac{dQ_{n,p}}{dt} &= \beta_{hh}Q^{h\to hh}_{n,p} + \gamma_h\delta_{n,0} - (\beta_{hh}+\gamma_h)Q_{n,p}\end{aligned} \tag{3.5}$$

The set of backward equations (1.13) can be solved using probability generating functions (PGFs) [11, 12]. Let $G_a(x,y,u,z,w;t)$ be the PGF for the joint distribution of the dynamic variables when a single animal host was infected at time 0:

$$G_a(x,y,u,z,w;t) = \sum_{i,j,k,l,m} P_{i,j,k,l,m}(t)\cdot x^i y^j u^k z^l w^m \tag{3.6}$$

Similarly, let $G_h(u,w;t)$ be the PGF for the joint distribution of $(I_h(t), Z_{h,s}(t))$ where a single human host is infected at time 0:

$$G_h(u,w;t) = \sum_{n,p} Q_{n,p}(t)\cdot u^n w^p \tag{3.7}$$

The PGF for $P^{a\to aa}_{i,j,k,l,m}(t)$ is simply G_a^2 reflecting the independence of branching process emanating out of two individuals. Similarly, the PGF for $P^{a\to ah}_{i,j,k,l,m}(t)$ is G_aG_h and that for $Q^{h\to hh}_{n,p}$ is G_h^2. Using the PGF representation we now write down the following backward equation:

$$\begin{aligned}\frac{\partial G_a}{\partial t} &= \beta_{aa}G_a^2 + \gamma_a y + \hat{\beta}_{ah}G_aG_hz - (\beta_{aa}+\hat{\beta}_{ah}+\gamma_a)G_a \\ \frac{\partial G_h}{\partial t} &= \beta_{hh}G_h^2 w + \gamma_h - (\beta_{hh}+\gamma_h)G_h\end{aligned} \tag{3.8a}$$

The initial conditions for the PDEs are:

$$\begin{aligned}G_a(x,y,u,z,w;0) &= x \\ G_h(u,w;0) &= u\end{aligned} \tag{3.8b}$$

which encode the information that there is a one infected animal host at time 0. The equation for G_h can be solved exactly. The solution is provided in [11, 12] and we reproduce it here:

$$G_h(u,w;t) = \frac{A_h(B_h-u)+B_h(u-A_h)e^{-\beta_{hh}w(B_h-A_h)t}}{(B_h-u)+(u-A_h)e^{-\beta_{hh}w(B_h-A_h)t}} \tag{3.9a}$$

where $A_h(w)$ and $B_h(w)$ are solutions of the following quadratic equation such that $0 < A_h < 1 < B_h$:

$$\mathcal{R}_0^{hh} w s^2 - (\mathcal{R}_0^{hh} + 1)s + 1 = 0 \tag{3.9b}$$

The PGF G_h quantifies the distribution of a single chain of infections that originates from a single primary infection. The more interesting aspect of the zoonoses dynamics is captured by the first equation (for G_a). A full analytical solution to this process has recently been reported [50], but extracting information specifically about the first spillover event into humans — which is our primary focus here — is complicated to derive from that general result. To address first passage time phenomena more directly, we introduce a simpler subset process below and derive its solution.

3.3.1 Analytical solution for a subset process

The distribution of $(I_a, R_a, Z_{h,p})$ is governed by a reduced set of reaction equations:

$$\begin{aligned}
(I_a, R_a, Z_{h,p}) &\xrightarrow{\beta_{aa} I_a} (I_a + 1, R_a, Z_{h,p}) \\
(I_a, R_a, Z_{h,p}) &\xrightarrow{\gamma_a I_a} (I_a - 1, R_a + 1, Z_{h,p}) \\
(I_a, R_a, Z_{h,p}) &\xrightarrow{\hat{\beta}_{ah} I_a} (I_a, R_a, Z_{h,p} + 1)
\end{aligned} \tag{3.10}$$

Let $\mathcal{G}_a(x,y,z;t)$ represent the PGF for the distribution of the above process. The PDE for the PGF is given by:

$$\frac{\partial \mathcal{G}_a}{\partial t} = \beta_{aa}\mathcal{G}_a^2 + \gamma_a y - (\beta_{aa} + \hat{\beta}_{ah} + \gamma_a)\mathcal{G}_a + \hat{\beta}_{ah}\mathcal{G}_a z \tag{3.11a}$$

with the initial condition

$$\mathcal{G}_a(x, y, z; 0) = x \tag{3.11b}$$

Following the methods outlined in [12], we obtain the following solution to the PDE:

$$\mathcal{G}_a(x,y,z;t) = \frac{A_a(B_a-x)+B_a(x-A_a)e^{-\beta_{aa}(B_a-A_a)t}}{(B_a-x)+(x-A_a)e^{-\beta_{aa}(B_a-A_a)t}} \tag{3.12a}$$

$A_a(y, z)$ and $B_a(y, z)$ are roots of the following quadratic equation such that $0 < A_a < 1 < B_a$.

$$\mathcal{R}_0^{aa} s^2 - \left(\mathcal{R}_0^{aa}+1+\mathcal{R}_0^{ah}(1-z)\right) s + y = 0 \tag{3.12b}$$

The distribution reported here has been solved before in the context of a human-only epidemic process with two types of hosts [51]. In subsequent sections, we shall require the value of roots at the point $z = 0$. Adopting notation from [52], we define:

$$\begin{aligned} V_0(y) &= A_a(y, 0)\,, \quad v_0 = A_a(1, 0) \\ V_1(y) &= B_a(y, 0)\,, \quad v_1 = B_a(1, 0) \end{aligned} \tag{3.12c}$$

3.3.2 First passage time and probability of spillover

We define the time to spillover as the first passage time T for human infection, i.e., as the time when the first primary infection occurs in the human hosts. The distribution of first passage times is given by:

$$\begin{aligned} \mathbb{P}[T \le t] &= \mathbb{P}[Z_{h,p}(t) > 0] \\ &= 1 - \mathcal{G}_a(1, 1, 0; t) \\ &= 1 - \frac{v_0(v_1-1)+v_1(1-v_0)e^{-\beta_{aa}(v_1-v_0)t}}{(v_1-1)+(1-v_0)e^{-\beta_{aa}(v_1-v_0)t}} \end{aligned} \tag{3.13}$$

This distribution is shown for various parameter values in Figure 3.2, along with results of discrete event simulation drawn from the underlying set of reactions. Simulations were done

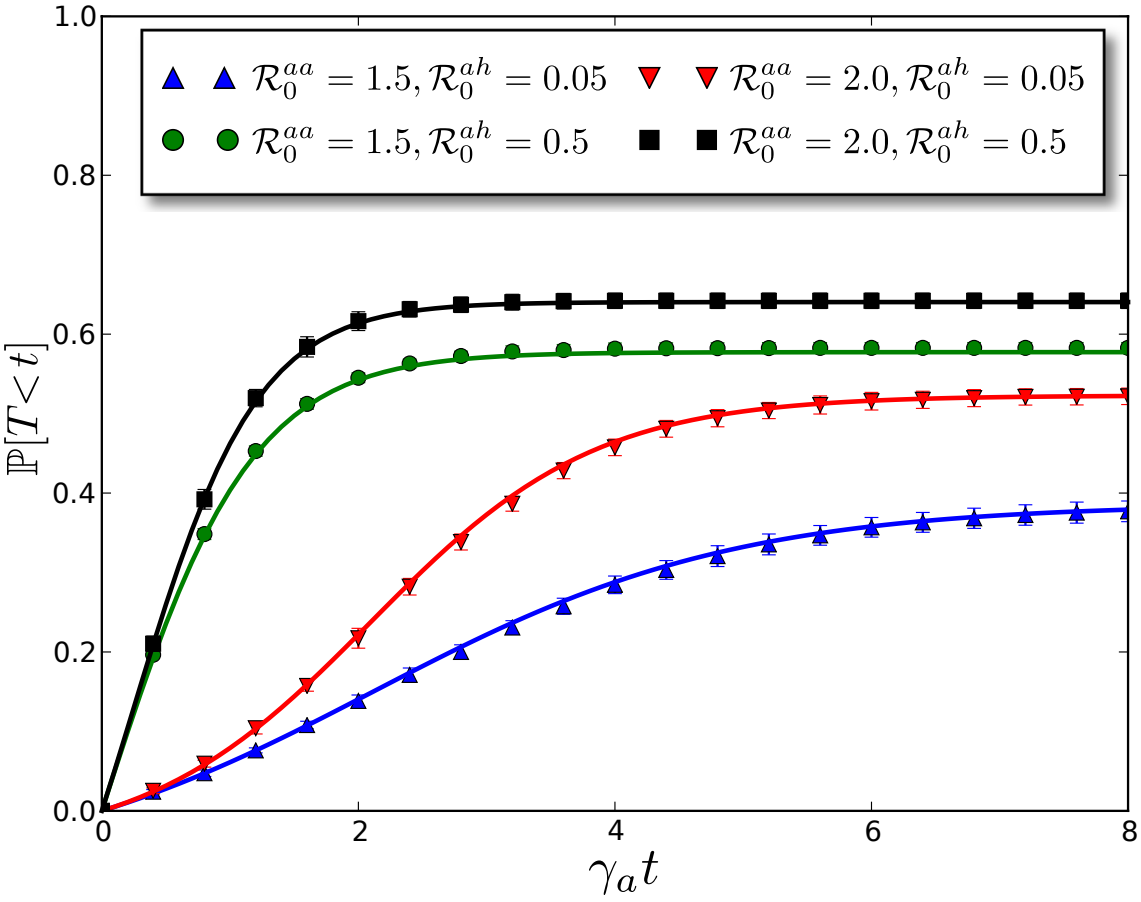


Figure 3.2: (Color online) First-passage time distribution for spillover into the human population $\mathbb{P}[T < t]$, comparing the analytical calculation given by eq. (3.13) (solid line) with the results of discrete event simulation (using Gillespie's direct method, for finite system size $N_a = N_h = 10^3$). The x-axis is time normalized by the mean infectious period ($1/\gamma_a$), of the animal species. The markers represent the mean of 8000 simulation runs (see appendix F).

using Gillespie's direct method [53] for reaction kinetics. It can be seen from Figure 3.2 that the distribution is defective since the disease can go extinct in the animal population before the primary transmission occurs in the human population. The distribution can be used to calculated moments of the first passage time conditioned on the occurrence of spillover (see Appendix D.1 for derivation):

$$\mathbb{E}[T^n \,|\, T < \infty] = \frac{n!}{(\beta_{aa})^n \,(v_0 - 1)\,(v_1 - v_0)^{n-1}} \mathrm{Li}_n\left(\frac{v_0 - 1}{v_1 - 1}\right) \tag{3.14}$$

where Li_n is the polylogarithm function of order n. The mean and the standard deviation of the first passage time are plotted as a function of $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{ah}$ in Figure 3.3. The conditional nature of the distribution leads to a non-monotonic dependence on $\mathcal{R}_0^{aa}$. First passage times for $\mathcal{R}_0^{aa} < 1$ are limited by the timescale for the eventual extinction in the animal population: spillover must occur quickly if it is going to happen at all. The expected time to extinction in the animal population diverges as $\mathcal{R}_0^{aa} \to 1$ leading to an increase in the mean first passage time. The mean also decreases with increasing $\mathcal{R}_0^{ah}$ because of the increasing rate of animal-to-human transmission.

The distribution of first passage times also gives us a way to calculate the probability of spillover as

$$\mathbb{P}[\text{spill}] = \mathbb{P}[T < \infty] = 1 - v_0 \tag{3.15}$$

Using the law of total probability, the probability of spillover can be decomposed as follows:

$$\begin{aligned} \mathbb{P}[\text{spill}] = {} & \mathbb{P}[\text{spill} \,|\, \text{small outbreak}] \cdot \mathbb{P}[\text{small outbreak}] \\ & + \mathbb{P}[\text{spill} \,|\, \text{large outbreak}] \cdot \mathbb{P}[\text{large outbreak}] \end{aligned} \tag{3.16}$$

where the conditioning is done on the state of the outbreak in the animal population. For $\mathcal{R}_0^{aa} \leq 1$, all animal outbreaks are small. For $\mathcal{R}_0^{aa} > 1$, the probability of a large animal outbreak is non-zero. In the infinite size limit, it is implicitly assumed that

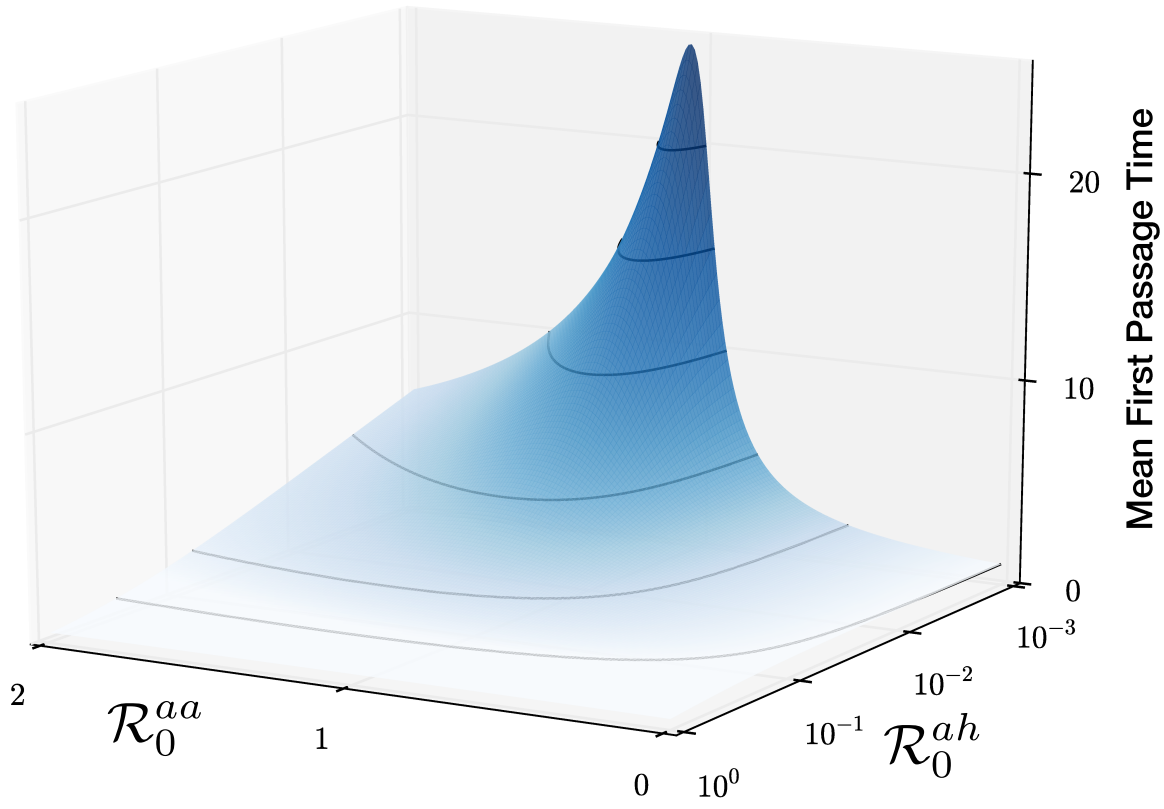


Figure 3.3: (Color online) The mean time to spillover (in units of the mean infectious period of animal hosts) as a function of $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{ah}$. Solid lines on the surface represent contours for the mean values. Gradient in the shade represents the standard deviation of the distribution (dark:high, light:low spanning the range $[0.4, 18.8]$ on a log scale).

$\mathbb{P}[\text{spill} \,|\, \text{large outbreak}] = 1$. Using this result and $\mathbb{P}[\text{large outbreak}] = 1 - 1/\mathcal{R}_0^{aa}$ in eq. (3.16), we can calculate $\mathbb{P}[\text{spill} \,|\, \text{small outbreak}]$ where $\mathcal{R}_0^{aa} > 1$, as follows:

$$\mathbb{P}[\text{spill} \,|\, \text{small outbreak}] = \begin{cases} 1 - \mathcal{R}_0^{aa} v_0 & \text{if } \mathcal{R}_0^{aa} \leq 1, \\ 1 - v_0 & \text{if } \mathcal{R}_0^{aa} > 1. \end{cases} \tag{3.17}$$

The probability of spillover as a function of relevant model parameters (eq. (3.15)) is shown in Figure 3.4 as the upper surface; also shown is the probability of spillover given that there is a small outbreak in the animal population (eq. (3.17)). We also calculate finite size corrections to the probability of spillover in Appendix D.2. The results of the calculation are shown in Figure 3.5.

Stochastic models offer a stark contrast to deterministic models that invariably associate spillover events with large outbreaks in the animal population. While large outbreaks do enhance the risk of spillover, small outbreaks also contribute as can be seen from Figure 3.4. This result indicates that some spillovers may be almost impossible to trace back in the animal population if they arise from a small outbreak where only a few animal hosts were infected and no contact tracing data is available.

3.3.3 Prevalence in the animal population at spillover

The distribution of infectious and recovered hosts in the animal population at the first passage time can be calculated by methods outlined in [52]. By interpreting the spillover process as a linear birth-death-killing (BDK) process, the distribution of infectious hosts at spillover is the same as the distribution of killing positions in the BDK process – geometrically distributed with parameter $1 - 1/v_1$ where v_1 was defined in eq. (3.12c). The calculation can be extended to include recovered hosts as well. The joint distribution of

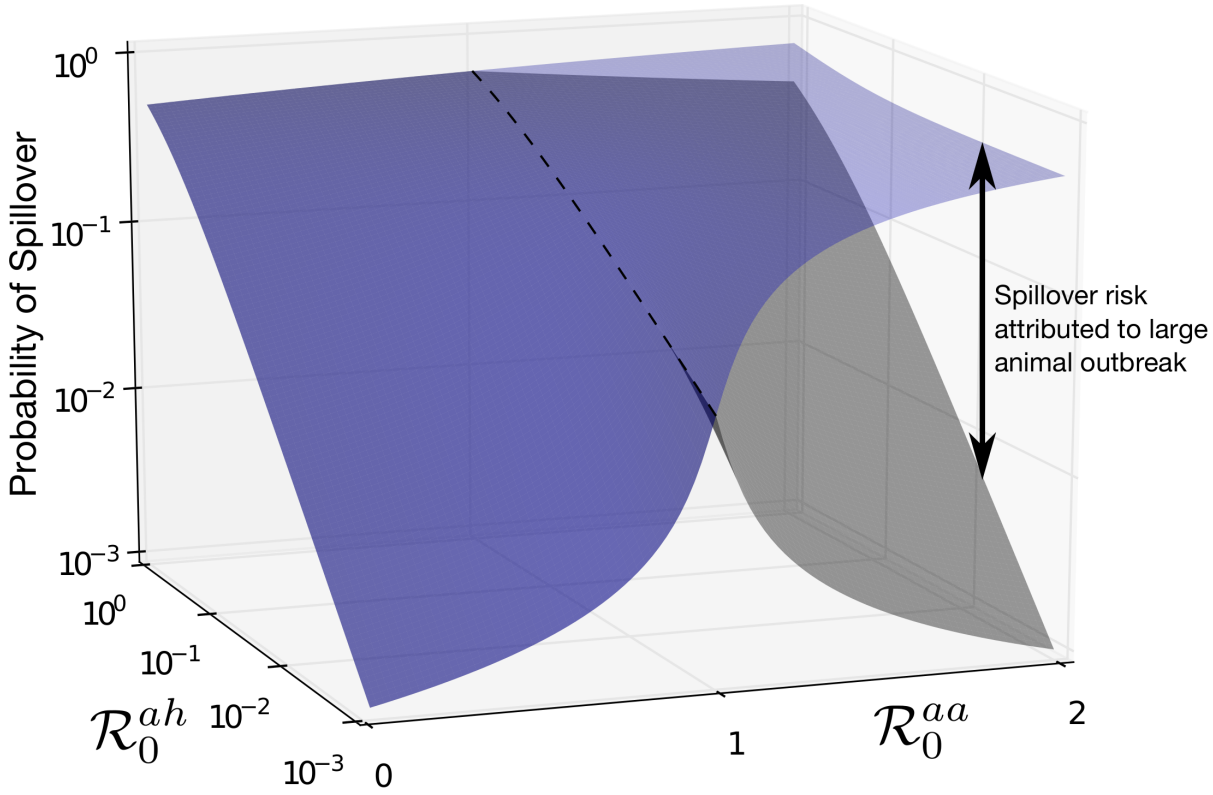


Figure 3.4: (Color online) Probability of spillover (blue, upper surface) and the conditional probability of spillover given a small outbreak in the animal population (gray,lower surface). The dashed line marks the separation between the two surfaces at $\mathcal{R}_0^{aa} = 1$. The difference between the two surfaces gives the contribution of large animal outbreaks to spillover risk.

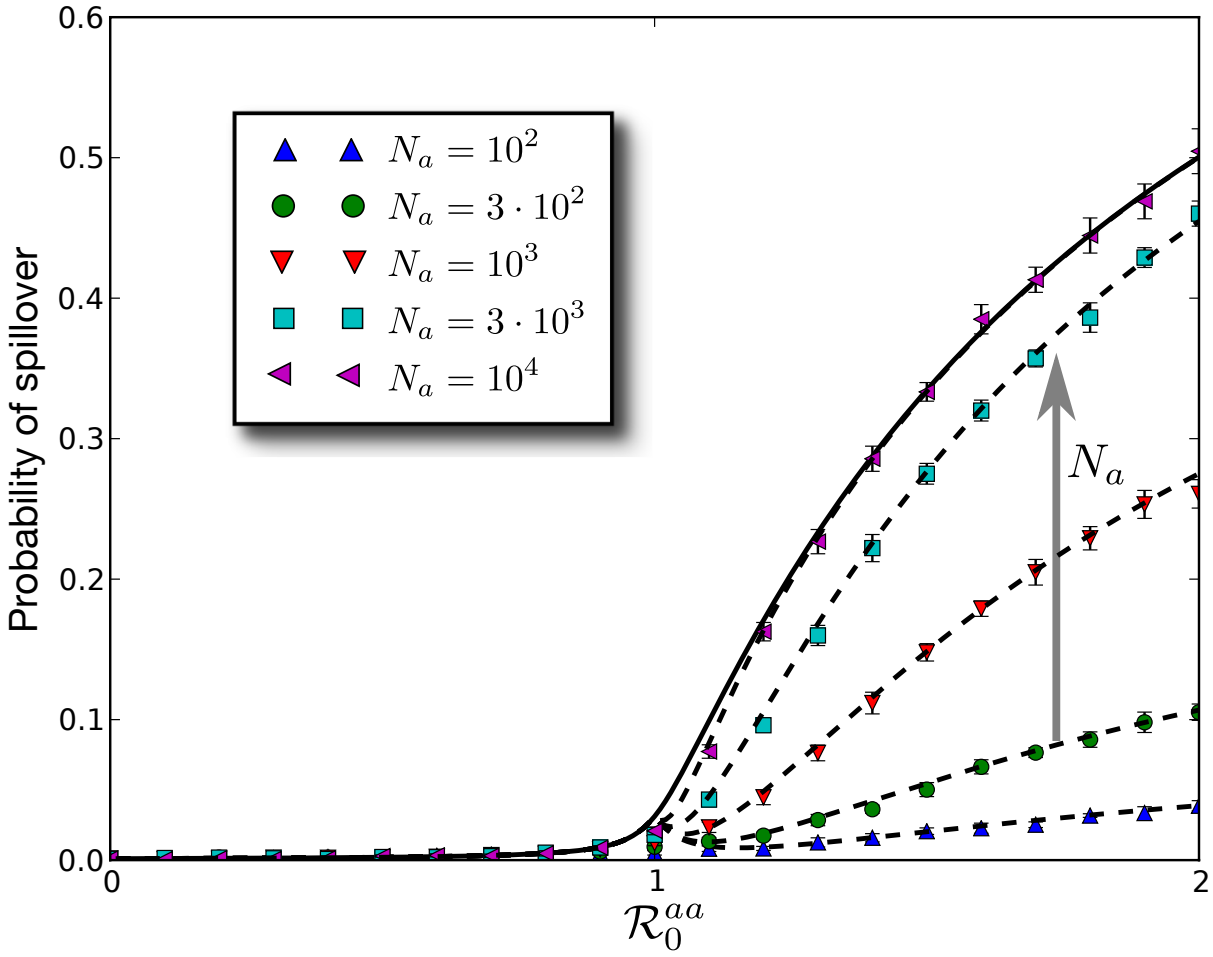


Figure 3.5: (Color online) Finite-size corrections to the probability of spillover. Dashed lines represent the analytical solution (eq. (D.6)) for different values of N_a. The solid line represents the solution from the linear birth-death process (eq. (3.15)). Colored markers represents values calculated from 10,000 simulation runs done using Gillespie's direct method (see appendix F). All results are for fixed $\mathcal{R}_0^{ah} = 10^{-3}$.

infectious and recovered animal hosts at first passage time is generated by the following PGF:

$$\begin{aligned} J(x,y) &= \frac{x\hat{\beta}_{ah}}{1-v_0}\int_0^\infty \frac{\partial \mathcal{G}_a(x,y,0;t)}{\partial x}dt \\ &= \frac{x\hat{\beta}_{ah}}{\beta_{aa}(1-v_0)\left(V_1(y)-x\right)} \\ &= \frac{x(v_1-1)}{V_1(y)-x} \end{aligned} \tag{3.18}$$

The simplification in the last step of eq. (3.18) comes from the fact that v_0 and v_1 are roots of eq. (3.12b), i.e.,

$$(v_1-1)(1-v_0) = \frac{\hat{\beta}_{ah}}{\beta_{aa}} = \frac{\mathcal{R}_0^{ah}}{\mathcal{R}_0^{aa}} \tag{3.19}$$

The mean number of infectious and recovered animal hosts at the first passage time T are given by:

$$\langle I_a(T)\rangle = \left.\frac{\partial J}{\partial x}\right|_{(1,1)}, \quad \langle R_a(T)\rangle = \left.\frac{\partial J}{\partial y}\right|_{(1,1)} \tag{3.20}$$

A plot of the mean number of infectious animal hosts at first passage time is shown as a function of $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{ah}$ in Figure 3.6. We can sample analytically the distributions represented in eq. (3.18), and compare with the results of stochastic simulations for finite system sizes, as is shown in Figure 3.7. As seen in the Figure 3.7 (a) for the expected number of infected animal hosts, the tail of the analytical distribution overestimates the prevalence slightly because of epidemic saturation that occurs in finite populations.

Given a prevalence of n infected animal hosts at spillover (and no information about recovered hosts), the maximum likelihood estimate for the parameters yields the equation $v_1 = n/(n-1)$. From eqs. (3.12b) and (3.12c) we arrive at the following relationship between the model parameters

$$\bar{R}_0^{aa} = (n-1)\left(\bar{R}_0^{ah} + n^{-1}\right) \tag{3.21}$$

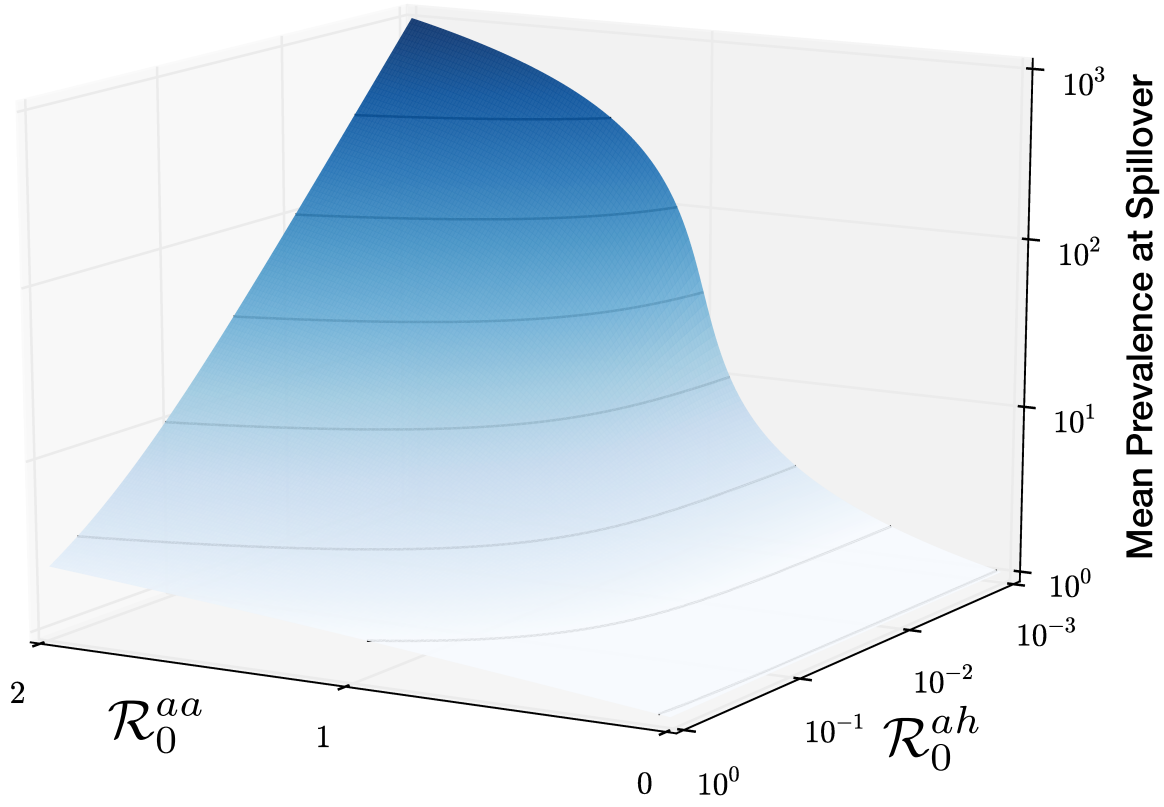


Figure 3.6: (Color online) The expected number of infectious animal hosts at the time of first primary human infection as a function of $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{ah}$. Solid lines on the surface represent contours for the mean values. Gradient in the shade represents the standard deviation of the distribution (dark:high, light:low spanning the range $[10^{-2}, 10^3]$ on a log scale).

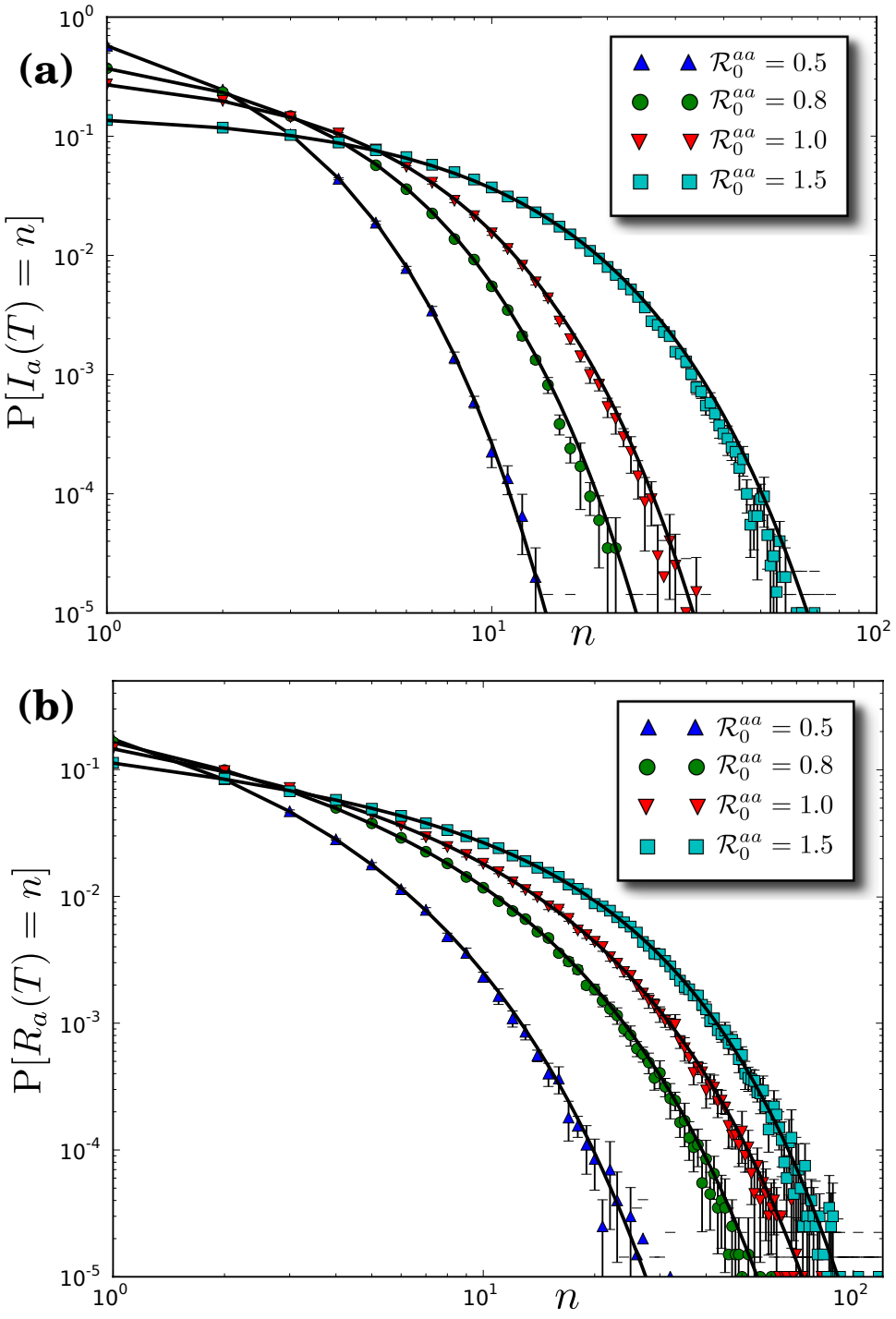


Figure 3.7: (Color online) The marginal distribution of the number of infectious animal hosts **(a)**, and the number of recovered animal hosts **(b)** at first passage time T for finite system size ($N_a = N_h = 1000$). Solid lines represents the analytical solution obtained by sampling from the PGF in eq. (3.18). Colored markers represents values calculated from 2×10^5 simulation runs done using Gillespie's direct method (see appendix F). All results are for fixed $\mathcal{R}_0^{ah} = 0.1$

.

where $\bar{R}_0^{aa}$ and $\bar{R}_0^{ah}$ are estimators for the corresponding parameters.

While the first passage time reveals the timescale of spillover, the disease prevalence reveals the state of the system at spillover; the different dependence on transmission parameters for these two quantities, however, has interesting implications for parameter estimation and disease control (compare Figures 3.3 and 3.6). Our results indicate that the fluctuations in the prevalence at spillover increase with $\mathcal{R}_0^{aa}$, in contrast to the first passage time which has the highest fluctuations near $\mathcal{R}_0^{aa} = 1$. In the absence of animal surveillance, the first spillover into humans is usually the point at which the disease is first detected and control interventions are initiated [49]. For $\mathcal{R}_0^{aa}$ substantially larger than 1, the spillover is likely to happen relatively early, but the disease prevalence may be quite large, making control difficult. In contrast, for $\mathcal{R}_0^{aa}$ close to 1, the disease is likely to be detected late, but there will be a low prevalence in the animal population at that time. This is encouraging for public health interventions aimed at controlling the disease in the animal population, although the long delay before detection introduces greater uncertainties as to whether other factors might need to be included in a more complicated model (such as demographic changes in the animal and human populations, or evolution of pathogen virulence).

3.4 Branching Processes

In the limit of $t \to \infty$, the SIR process is isomorphic to a Galton-Watson branching process where the offspring are the new infections produced by an infected host. The distribution of outbreak sizes can be calculated by making the *tree-like* approximation for small outbreaks.

3.4.1 Distribution of outbreak sizes

For the animal population, let $H_a(z)$ be the PGF for the distribution of outbreak sizes. From eq. (3.12a), we obtain:

$$\begin{aligned} H_a(z) &= \mathcal{G}_a(1, z, 1; \infty) = A_a(z, 1) \\ &= \frac{\mathcal{R}_0^{aa} + 1 - \sqrt{(\mathcal{R}_0^{aa} + 1)^2 - 4\mathcal{R}_0^{aa} z}}{2\mathcal{R}_0^{aa}} \end{aligned} \tag{3.22}$$

Let $H_{h,p}(x)$ be the PGF for the distribution of primary infections in the human population. Then, from eq. (3.12a), we obtain:

$$\begin{aligned} H_{h,p}(x) &= \mathcal{G}_a(1, 1, x; \infty) = A_a(1, x) \\ &= \frac{\mathcal{R}_0^{aa}+1+\mathcal{R}_0^{ah}(1-x)-\sqrt{(\mathcal{R}_0^{aa}+1+\mathcal{R}_0^{ah}(1-x))^2-4\mathcal{R}_0^{aa}}}{2\mathcal{R}_0^{aa}} \end{aligned} \tag{3.23}$$

Each primary infected host in the human population acts as the progenitor for a branching process comprising of secondary infections. Let $\hat{H}_{h,s}(x)$ be the PGF for the distribution of secondary infections emanating from a primary progenitor. Then, from eq. (3.9a) we obtain:

$$\begin{aligned} \hat{H}_{h,s}(z) &= G_h(1, z; \infty) = A_h(z) \\ &= \frac{\mathcal{R}_0^{hh} + 1 - \sqrt{(\mathcal{R}_0^{hh} + 1)^2 - 4\mathcal{R}_0^{hh} z}}{2\mathcal{R}_0^{hh} z} \end{aligned} \tag{3.24}$$

The PGF for the joint distribution of primary and secondary infections can be written as:

$$H_h(x, z) = H_{h,p}(x\hat{H}_{h,s}(z)) \tag{3.25}$$

The PGF for the total number (irrespective of whether the infection was primary or secondary) is given by:

$$H_h(z) = H_{h,p}(zH_{h,s}(z)) \tag{3.26}$$

The probability of an outbreak being small is simply $H_h(1)$. This probability is shown as a function of $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{hh}$, for fixed $\mathcal{R}_0^{ah}$, in Figure 3.8A. Lastly, the PGF for secondary infections is given by:

$$H_{h,s}(z) = H_h(1, z) \tag{3.27}$$

Following [54], we can extract the probability of n human hosts getting infected using the Cauchy integral formula:

$$\mathbb{P}[Z_h(\infty) = n] = \frac{1}{2\pi i} \oint \frac{H_h(z)}{z^{n+1}}\, dz \tag{3.28}$$

where the integral is done over the unit circle $|z| = 1$ in the complex plane. Similarly, the joint probability distribution can be extracted by extending the Cauchy integral formula to higher dimensions:

$$\mathbb{P}[Z_{h,p}(\infty) = m, Z_{h,s}(\infty) = n] = \frac{1}{(2\pi i)^2} \oint \oint \frac{H_h(x, z)}{x^{m+1} z^{n+1}}\, dx\, dz \tag{3.29}$$

where the integrals are over two unit circles in the x and z complex planes.

3.4.2 Critical threshold

The critical threshold is defined as the point in parameter space where the average outbreak size diverges [39, 40] and the probability of a large outbreak becomes greater than 0. We introduce the following notation for the average outbreak sizes.

$$\begin{aligned} \langle n \rangle_a &\equiv \mathbb{E}[R_a(\infty)] \\ \langle n \rangle_h &\equiv \mathbb{E}[R_h(\infty)] \end{aligned} \tag{3.30}$$

For the animal population,

$$\begin{aligned} \langle n \rangle_a &= H_a'(1) \\ &= \frac{1}{1 - \mathcal{R}_0^{aa}} \end{aligned} \tag{3.31}$$

which yields the condition $\mathcal{R}_0^{aa} = 1$ as the critical threshold. For the human population,

$$\begin{aligned}\langle n \rangle_h &= H_h'(1) \\ &= H_{h,p}'(1)\left\{1 + H_{h,s}'(1)\right\} \\ &= \frac{\mathcal{R}_0^{ah}}{(1-\mathcal{R}_0^{aa})(1-\mathcal{R}_0^{hh})}\end{aligned} \tag{3.32}$$

From the above expression, the critical threshold for the human population is given by $\max(\mathcal{R}_0^{aa}, \mathcal{R}_0^{hh}) = 1$. Thus, large outbreaks in the human population are possible even if $\mathcal{R}_0^{hh} < 1$, emphasizing the potential importance of *spillover-driven large outbreaks*.

3.4.3 Asymptotic scaling near the critical threshold

The scaling of the outbreak sizes near the threshold boundary can be investigated through the singularity analysis of the associated generated function (see Appendix D.3 for details). The threshold boundary can be divided into three parts: (1) $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} < 1$, (2) $\mathcal{R}_0^{aa} < 1, \mathcal{R}_0^{hh} = 1$ and (3) $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} = 1$. (Subscripts 1,2 and 3 below refer to these three parts of the boundary, respectively.) Along the two lines of the threshold boundary excluding the multi-critical point, the distribution of outbreak sizes scale as:

$$P_i(n) \sim \zeta_i^{-n}\, n^{-3/2}, \quad i = 1, 2 \tag{3.33}$$

whereas near the multi-critical point the distribution scales as:

$$P_3(n) \sim \zeta_3^{-n} n^{-5/4} \tag{3.34}$$

The variables ζ_i are a function of the distance from the threshold boundary. The ζ's are defined in terms of $\Delta_a = 1 - \mathcal{R}_0^{aa}$ and $\Delta_h = 1 - \mathcal{R}_0^{hh}$.

$$\begin{aligned}
\zeta_1 &= 1 + \frac{\Delta_h \Delta_a^2}{4\mathcal{R}_0^{ah}} + \mathcal{O}(\Delta_h \Delta_a^3) \\
\zeta_2 &= 1 + \frac{\Delta_h^2}{4} \\
\zeta_3 &= 1 + \frac{\Delta_h^2}{4} \quad \text{where} \;\; \Delta_h = \frac{\Delta_a^2}{2\mathcal{R}_0^{ah}} + \mathcal{O}(\Delta_a^3)
\end{aligned} \tag{3.35}$$

The scaling laws with exponential cutoffs are shown in Figure 3.8 along different crossings of the threshold boundary. Note that ζ_3 is a valid exponential cutoff only on a parabolic curve near the multi-critical point. The problem of estimating the corrections away from this curve is non-trivial and rigorous results are still an open problem. But intuitively we know that in this case, the generating function will have two singularities which are coalescing at the multi-critical point. In such a scenario, there will be a crossover regime where the power-law exponent switches from 3/2 to 5/4 depending on the distance from the threshold boundary.

The appearance of a different scaling exponent at the multi-critical point points to a different universality class emerging from the simultaneous divergence of two individual SIR processes in our system. We note that the same $n^{-5/4}$ scaling – arising from one critical process driving another – has been reported recently in a different, albeit related, multitype critical branching process intended to model multistage SIR infections [15].

3.4.4 Finite size scaling at critical threshold

Using the heuristic arguments presented in [15], we can calculate how the average outbreak size scales with system size at the threshold boundary $\max(\mathcal{R}_0^{aa}, \mathcal{R}_0^{hh}) = 1$. Let M_a be the 'maximal' size of an outbreak in the animal population, when $\mathcal{R}_0^{aa} = 1$, such that an

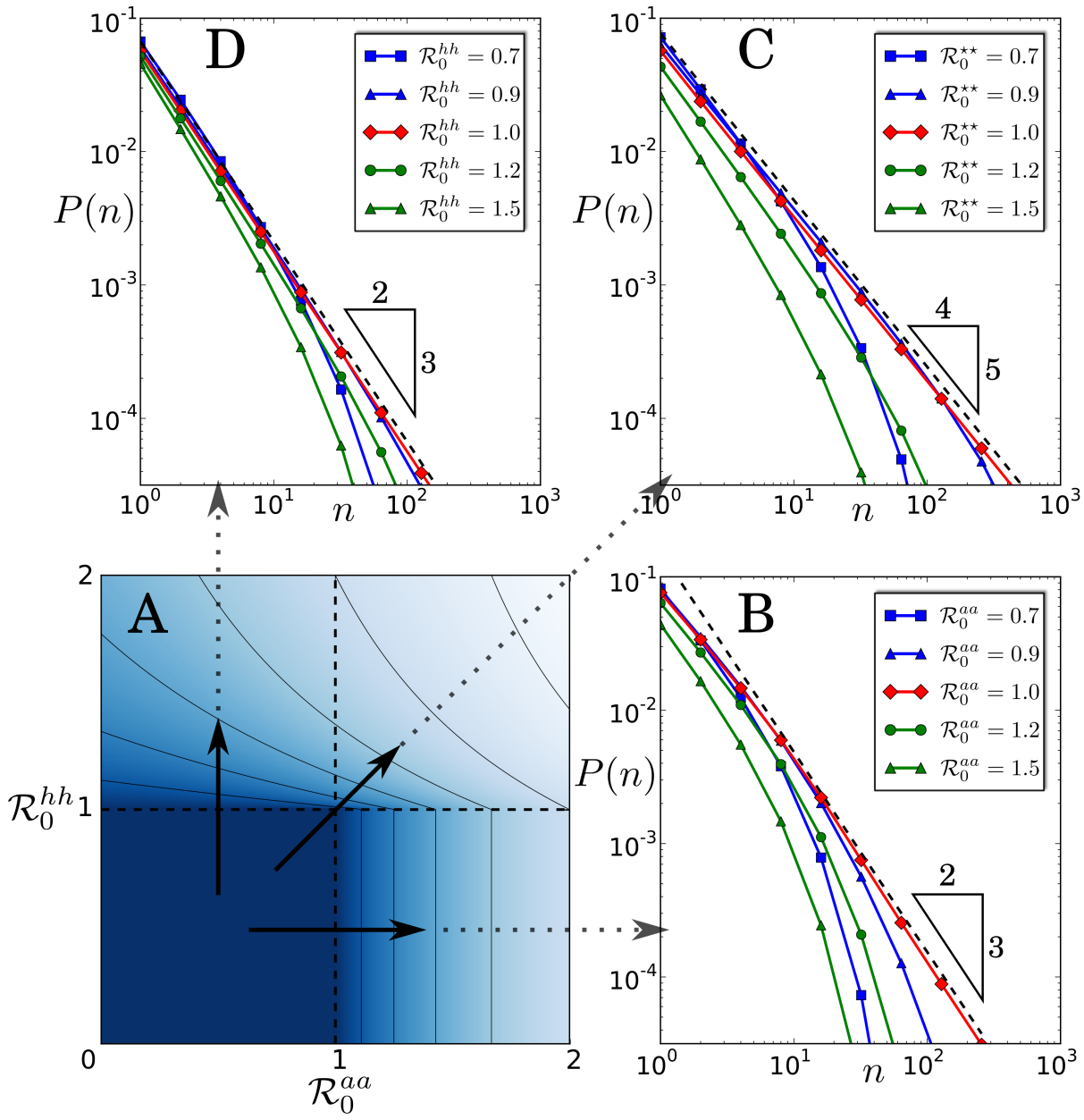


Figure 3.8: (Color online) The distribution of sizes of small human outbreaks. (A) Colormap for the probability that an outbreak in the human hosts is small spanning the range [0.36 (light), 1.0 (dark)]. Solid lines represent constant-probability contours in the colormap. (B-D) Probability of having a small outbreak of size n at different crossings of the threshold boundary. All results are for fixed $\mathcal{R}_0^{ah} = 0.1$

.

outbreak cannot exceed this size due to depletion of susceptible hosts [15]. The effective $\mathcal{R}_0^{aa}$ for a finite sized system reduces to:

$$\hat{\mathcal{R}}_0^{aa} = 1 - M_a/N_a \tag{3.36}$$

Using eq. (3.31), we obtain the following estimate for the scale of the average outbreak size:

$$\langle n \rangle_a \sim N_a/M_a \tag{3.37}$$

From the 3/2 scaling law for single-type SIR [15], we obtain a second estimate for the average outbreak size:

$$\langle n \rangle_a = \sum_{n=1}^{M_a} n \cdot n^{-3/2} \sim \sqrt{M_a} \tag{3.38}$$

Equating the two estimates and imposing self-consistency, one obtains the following scaling laws (see [15])

$$M_a \sim N_a^{2/3}, \quad \langle n \rangle_a \sim N_a^{1/3} \tag{3.39}$$

One can also calculate the scaling window that represents the distance from the threshold boundary within which the scaling law will hold [16]. For the animal SIR, the scaling window is given by:

$$|1 - \mathcal{R}_0^{aa}| \sim N_a^{-1/3} \tag{3.40}$$

The calculation for human outbreaks is separated into 3 cases (as highlighted in Figure 3.8 B,C,D). For $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} < 1$, the average outbreak size is given by substituting $\hat{\mathcal{R}}_0^{aa}$ in eq. (3.32), as follows:

$$\langle n \rangle_h \sim N_a/M_a = N_a^{1/3} \tag{3.41}$$

The second estimate is obtained by using the scaling law of 3/2 derived in eq. (D.15b).

$$\langle n \rangle_h = \sum_{n=1}^{M_h} n^{-1/2} \sim \sqrt{M_h} \tag{3.42}$$

Equating the two estimates reveals $M_h \sim N_a^{2/3}$. If $\mathcal{O}(N_a) \gg \mathcal{O}(N_h^{3/2})$, the scaling relation leads to the maximal outbreak exceeding the system size, which is physically inconsistent. Thus, the maximal outbreak scale needs to be capped at N_h, i.e.,

$$M_h \sim \min(N_a^{2/3}, N_h) \tag{3.43a}$$

From (3.43a), we can estimate that the crossover regime between the two scales in the min function is given by $N_h \sim N_a^{2/3}$. The scaling of average outbreak size is given by $\sqrt{M_h}$, i.e.,

$$\langle n \rangle_h \sim \min(N_a^{1/3}, N_h^{1/2}) \tag{3.43b}$$

The analytical result of eq. (3.43b) is validated in Figure 3.9. The scaling window near the boundary $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} < 1$ is given by:

$$|1 - \mathcal{R}_0^{aa}| \sim \max(N_a^{-1/3}, N_h^{-1/2}) \tag{3.44}$$

The case of $\mathcal{R}_0^{aa} < 1, \mathcal{R}_0^{hh} = 1$ results in the same calculations as for a single-type SIR. Thus, the scaling laws are the same as in eq. (3.39):

$$M_h \sim N_h^{2/3}, \quad \langle n \rangle_h \sim N_h^{1/3} \tag{3.45}$$

with the scaling window same as that for the simple SIR:

$$|1 - \mathcal{R}_0^{hh}| \sim N_h^{-1/3} \tag{3.46}$$

At the multicritical point, the effective basic reproduction numbers are:

$$\hat{\mathcal{R}}_0^{aa} = 1 - M_a/N_a, \quad \hat{\mathcal{R}}_0^{hh} = 1 - M_h/N_h$$

From (3.32), we arrive at the first estimate:

$$\langle n \rangle_h \sim \frac{N_a}{M_a} \frac{N_h}{M_h} = \frac{N_a^{1/3} N_h}{M_h} \tag{3.47}$$

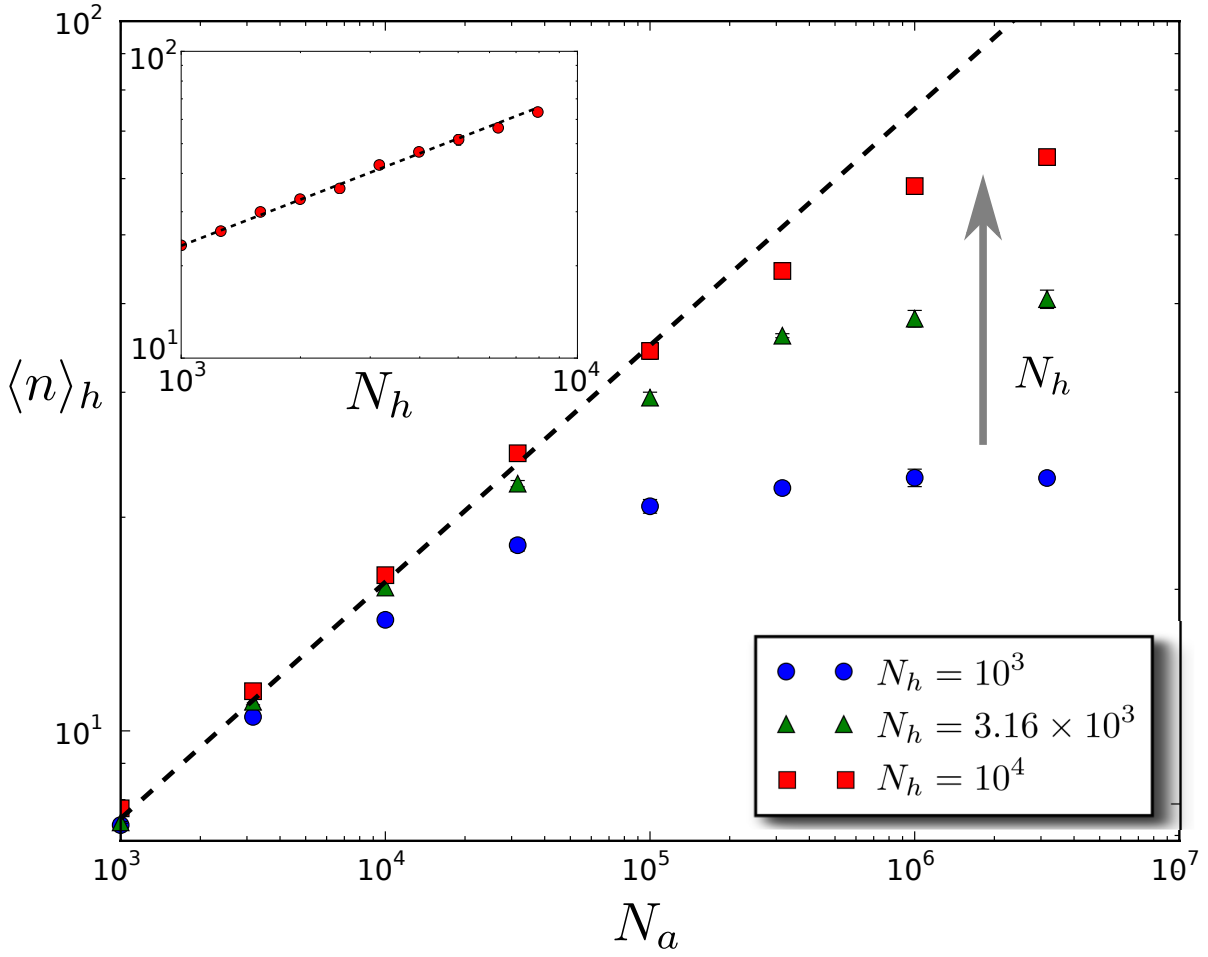


Figure 3.9: (Color online) Finite size scaling at the threshold boundary $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} < 1$. The plot shows the scaling law for average outbreak size in humans $\langle n \rangle_h \sim N_a^{1/3}$ and crossover to $N_h^{1/2}$ when $N_h \sim N_a^{2/3}$ on a log-log plot. The points are the average of 7×10^4 stochastic realizations. The dashed line has slope 1/3. (*Inset*) The average outbreak size $\langle n \rangle_h$ plotted against N_h on a log-log scale for fixed $N_a = 10^7$. The dashed line has slope of 1/2. The points are the average over 10^5 stochastic realizations (see appendix F). All results for $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{ah} = 0.5, \mathcal{R}_0^{hh} = 0.1$.

The second estimate is derived from eq. (D.21a):

$$\langle n \rangle_h = \sum_{n=1}^{M_h} n^{-1/4} \sim M_h^{3/4} \tag{3.48}$$

Equating the two estimates provides the scaling for the maximal outbreak size:

$$M_h \sim \left(N_a^{1/3} N_h\right)^{4/7} \tag{3.49}$$

Since the maximal outbreak can not exceed the system size:

$$M_h \sim \min\left(N_h, \left(N_a^{1/3} N_h\right)^{4/7}\right) \tag{3.50a}$$

The scale of the average outbreak size is given by:

$$\langle n \rangle_h \sim \min\left(N_h^{3/4}, \left(N_a N_h^3\right)^{1/7}\right) \tag{3.50b}$$

The crossover region in the multicritical case is $N_h \sim N_a^{4/9}$. The scaling window in this case would depend on both $\mathcal{R}_0^{aa}$ and $\mathcal{R}_0^{hh}$, i.e.,

$$|(1 - \mathcal{R}_0^{aa})(1 - \mathcal{R}_0^{hh})| \sim \max\left(N_h^{-3/4}, \left(N_a N_h^3\right)^{-1/7}\right) \tag{3.51}$$

The finite size scaling laws have important implications for determining whether a critical outbreak is spillover driven or intrinsically driven. Note that while the two lines of the threshold boundary have the same scaling in the distribution of outbreak sizes, the average outbreak size scales differently along those lines. Whereas an intrinsically driven critical outbreak ($\mathcal{R}_0^{aa} < 1, \mathcal{R}_0^{hh} = 1$) scales only with the abundance of the human host population, a spillover driven critical outbreak ($\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} < 1$) can depend on the abundance of the animal population. In addition, for $N_h \ll N_a^{2/3}$, a spillover-driven outbreak has a greater extent of $\mathcal{O}(N_h^{1/2})$ as compared to an intrinsically driven outbreak (simple SIR) which is capped at $\mathcal{O}(N_h^{1/3})$.

3.4.5 Probability of large outbreak

For the animal population, the probability of large outbreak is calculated as:

$$\begin{aligned}\mathbb{P}[R_a(\infty) = \infty] &= 1 - H_a(1) \\ &= 1 - \frac{1}{\mathcal{R}_0^{aa}}\end{aligned} \tag{3.52}$$

Similarly, the probability of a large human outbreak is calculated as follows. Assuming $\mathcal{R}_0^{ah} > 0$,

$$\begin{aligned}\mathbb{P}[R_h(\infty) = \infty] &= 1 - H_h(1) \\ &= 1 - H_{h,p}(1, H_{h,s}(1)) \\ &= \begin{cases} 0 & \text{if } \mathcal{R}_0^{hh} \leq 1 \text{ and } \mathcal{R}_0^{aa} \leq 1, \\ 1 - \dfrac{1}{\mathcal{R}_0^{aa}} & \text{if } \mathcal{R}_0^{hh} \leq 1 \text{ and } \mathcal{R}_0^{aa} > 1, \\ 1 - A_a\left(1, \dfrac{1}{\mathcal{R}_0^{hh}}\right) & \text{if } \mathcal{R}_0^{hh} > 1. \end{cases}\end{aligned} \tag{3.53}$$

The probability is shown as a surface plot in Figure 3.10.

If $\mathcal{R}_0^{hh} \leq 1$, an outbreak in the human population can be large if and only if the outbreak in the animal population is large. In such a case, the probability of a large human outbreak is equal to the probability of a large outbreak in the animal population, which is a function of only $\mathcal{R}_0^{aa}$ (see Figure 3.12). On the other hand, if $\mathcal{R}_0^{hh} > 1$, a large human outbreak can occur even if the animal outbreak is small. Figures 3.11 and 3.12 compare the analytical results with results from stochastic simulation. Away from the phase transitions at $\mathcal{R}_0^{aa} = 1$ and $\mathcal{R}_0^{hh} = 1$, the results from stochastic simulations show good agreement with the theory. Near the phase transition, the simulation results should converge to the theory for increasing N. Since the definition of a large outbreak becomes precise only in the limit of large system size, there are no finite size corrections that can be derived in this case.

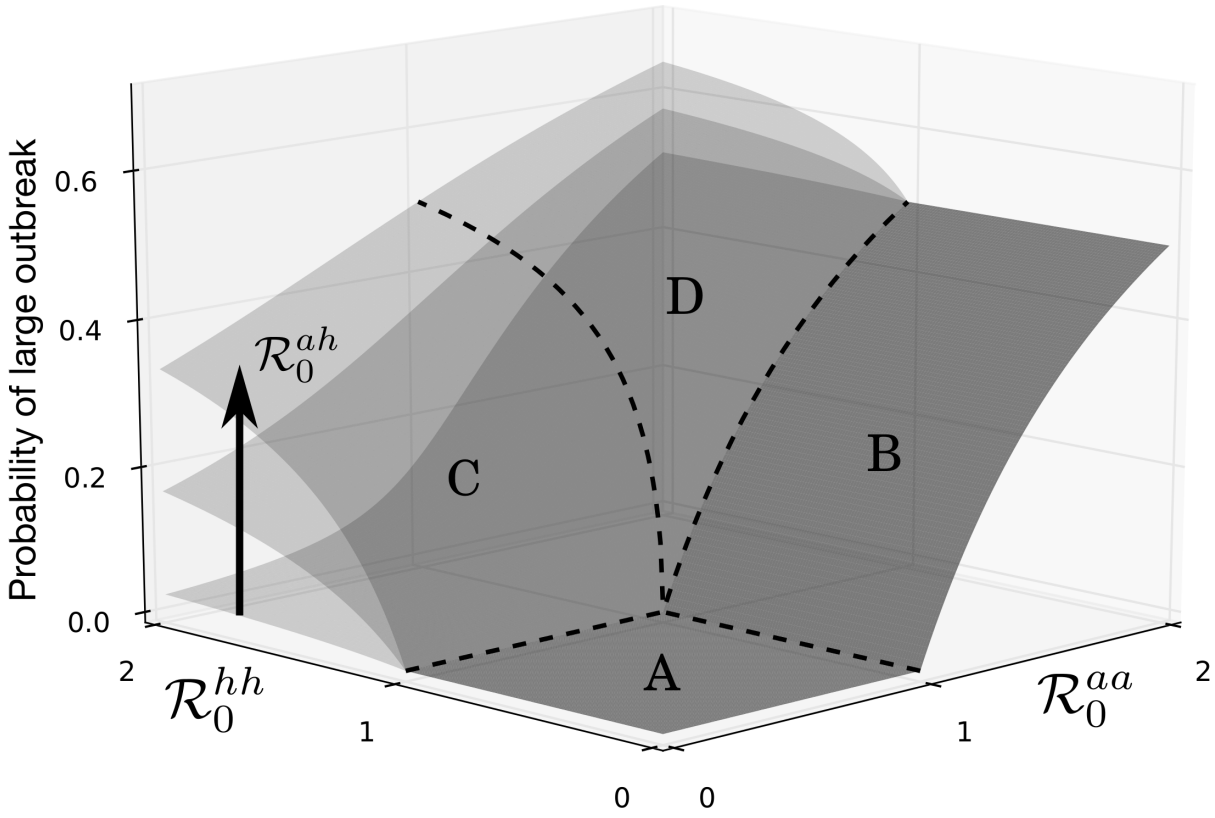


Figure 3.10: (Color online) Probability of a large outbreak in humans for increasing values of $\mathcal{R}_0^{ah} = [0.05, 0.4, 1.0]$. The upper surface is partitioned into 4 sections: (A) where all outbreaks are small, (B) where spillover-driven large outbreaks are possible, (C) where large outbreaks can only be sustained by human-to-human transmission, and (D) where sustained spillover and human-to-human transmission result in a large outbreak.

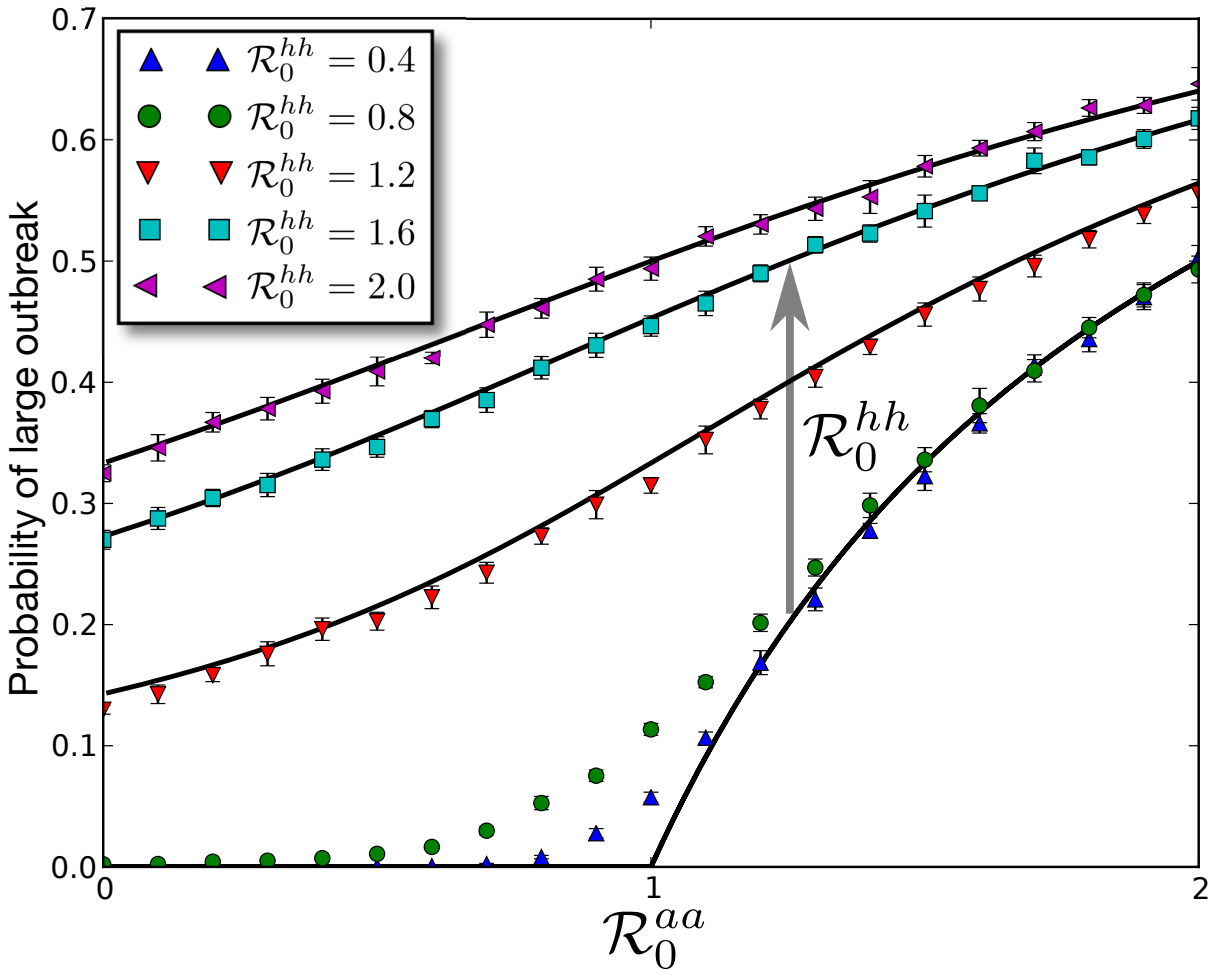


Figure 3.11: (Color online) The probability of a large human outbreak for finite populations ($N_a = N_h = 10^3$). The criteria for a large outbreak was chosen as 100 or more infected human hosts. The points represent the result of 10,000 stochastic simulations (see appendix F). The solid lines represent the analytical solution from eq. (3.53). The simulations do not agree with the analytical solution near the phase transition because of the chosen criteria for large outbreaks and finite size effects. All results are for fixed $\mathcal{R}_0^{ah} = 1$.

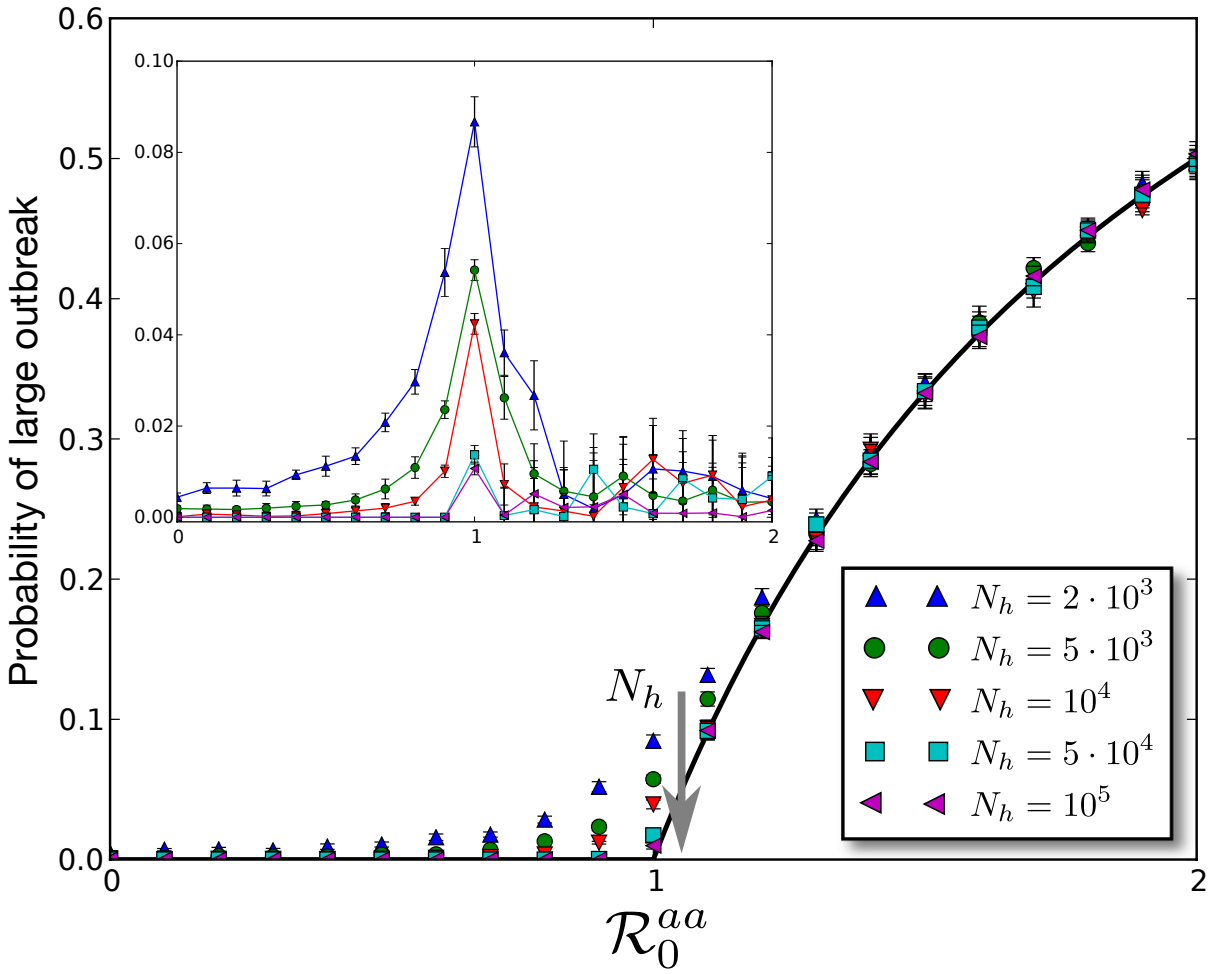


Figure 3.12: (Color online) The probability of a large human outbreak for $\mathcal{R}_0^{hh} = 0.8$ and varying N_h. The criteria for a large outbreak was chosen as the number of infected hosts being greater than 1% of the total population. The points represent the result of 10,000 stochastic simulations (see appendix F). The solid line is the analytical solution $\max(0, 1 - 1/\mathcal{R}_0^{aa})$. (*Inset*) The absolute difference between the analytical solution and finite size resuls. All results are for fixed $\mathcal{R}_0^{ah} = 0.1$.

3.5 Large outbreaks

The size of a large outbreak scales with the system size in the large population limit. The fraction of infected hosts can be calculated in several ways: (1) analytically solving the equivalent deterministic system, (2) hazard function [26] and (3) bond percolation on a complete graph [39, 55]. We use the hazard function to obtain the solution. First we write down the deterministic equations for our model.

3.5.1 Deterministic Equations

The deterministic representation of the model can be summarized through the following system of ODEs.

$$\begin{aligned}
\frac{d\mathcal{S}_a}{d\tau} &= -R_0^{aa}\mathcal{S}_a\mathcal{I}_a \\
\frac{d\mathcal{I}_a}{d\tau} &= R_0^{aa}\mathcal{S}_a\mathcal{I}_a - \mathcal{I}_a \\
\frac{d\mathcal{R}_a}{d\tau} &= \mathcal{I}_a
\end{aligned}$$

$$\begin{aligned}
\frac{d\mathcal{S}_{h,1}}{d\tau} &= -\lambda\,\mathcal{S}_{h,1}\mathcal{I}_a - \kappa\,R_0^{hh}\mathcal{S}_{h,1}(\mathcal{I}_{h,1,p}+\mathcal{I}_{h,1,s}+\mathcal{I}_{h,2}) \\
\frac{d\mathcal{I}_{h,1,p}}{d\tau} &= \lambda\,\mathcal{S}_{h,1}\mathcal{I}_a - \kappa\,\mathcal{I}_{h,1,p} \\
\frac{d\mathcal{I}_{h,1,s}}{d\tau} &= \kappa\,R_0^{hh}\mathcal{S}_{h,1}(\mathcal{I}_{h,1,p} + \mathcal{I}_{1,s} + \mathcal{I}_{h,2}) - \kappa\,\mathcal{I}_{h,1,s} \\
\frac{d\mathcal{R}_{h,1,p}}{d\tau} &= \kappa\,\mathcal{I}_{h,1,p} \\
\frac{d\mathcal{R}_{h,1,s}}{d\tau} &= \kappa\,\mathcal{I}_{h,1,s}
\end{aligned} \tag{3.54}$$

$$\frac{d\mathcal{S}_{h,2}}{d\tau} = -\kappa\, R_0^{hh}\mathcal{S}_{h,2}(\mathcal{I}_{h,1,p} + \mathcal{I}_{1,s} + \mathcal{I}_{h,2})$$
$$\frac{d\mathcal{I}_{h,2}}{d\tau} = \kappa\, R_0^{hh}\mathcal{S}_{h,2}(\mathcal{I}_{h,1,p} + \mathcal{I}_{1,s} + \mathcal{I}_{h,2}) - \kappa\,\mathcal{I}_{h,2}$$
$$\frac{d\mathcal{R}_{h,2}}{d\tau} = \kappa\,\mathcal{I}_{h,2}$$

where the variables $\mathcal{S}_\star, \mathcal{I}_\star, \mathcal{R}_\star$ are non-dimensional state variables that have been normalized by the total population of the species. Here, all dynamical variables for the human population are normalized by N_h and time is normalized by the average infectious period of the animal hosts. Two new variables are introduced here

$$\lambda = \frac{\rho R_0^{ah}}{\nu}, \quad \kappa = \frac{\gamma_h}{\gamma_a} \tag{3.55}$$

The non-dimensional parameters governing the dynamics of the system are: $(R_0^{aa}, \lambda, R_0^{hh}, \kappa)$. The initial conditions that we use to solve this system are given below

$$\mathcal{S}_a(0) = 1 - \frac{1}{N_a},\ \mathcal{I}_a(0) = \frac{1}{N_a},\ \mathcal{R}_a(0) = 0$$

$$\mathcal{S}_{h,1}(0) = \nu,\ \mathcal{S}_{h,2}(0) = 1 - \nu,\ \mathcal{I}_{h,\star}(0) = 0,\ \mathcal{R}_{h,\star}(0) = 0$$

3.5.2 Mean final size

Let $f_\star$ be the relative size of the infected hosts in the various host compartments in the limit of large system size for the stochastic version of the model.

$$\begin{aligned}
f_a &= \lim_{N_a\to\infty} \frac{\mathbb{E}[R_a(\infty)]}{N_a} \\
f_{h,p} &= \lim_{N_h\to\infty} \frac{\mathbb{E}[R_{h,1,p}(\infty)]}{N_h} \\
f_{h,s} &= \lim_{N_h\to\infty} \frac{\mathbb{E}[R_{h,1,s}(\infty)] + \mathbb{E}[R_{h,2}(\infty)]}{N_h} \\
f_{h,1} &= \lim_{N_h\to\infty} \frac{\mathbb{E}[R_{h,1,p}(\infty)] + \mathbb{E}[R_{h,1,s}(\infty)]}{N_h} \\
f_{h,2} &= \lim_{N_h\to\infty} \frac{\mathbb{E}[R_{h,2}(\infty)]}{N_h} \\
f_h &= f_{h,1} + f_{h,2} = f_{h,p} + f_{h,s}
\end{aligned} \tag{3.56}$$

Using survival analysis described in [26], we proceed with calculations for the various $f_\star$. The calculation is based on the result that in the limit of large system size the final epidemic size is the same as that given by solving the deterministic system of equations, i.e.,

$$f_\star = \mathcal{R}_\star(\infty) \tag{3.57}$$

For a randomly chosen susceptible host in the animal population, the cumulative hazard function is the probability of not getting infected before time t. This function can be calculated as follows

$$\Lambda_{aa}(t) = e^{-\int_0^t \beta_{aa}\mathcal{I}_a ds} \tag{3.58}$$

At steady state, the hazard function simplifies as follows

$$\Lambda_{aa}(\infty) = e^{-R_0^{aa}\mathcal{R}_a(\infty)} = e^{-R_0^{aa} f_a} \tag{3.59}$$

The probability of escaping infection would be $1 - f_a$. Equating this with equation (3.59), we obtain

$$1 - f_a = e^{-R_0^{aa} f_a} \tag{3.60}$$

Similarly for the human hosts, we first look at type 1 hosts (who are at risk of both primary and secondary transmissions). The hazard functions for the animal to human and human to human transmissions are given by

$$\begin{aligned} \Lambda_{ah}(\infty) &= e^{-\lambda f_a} \\ \Lambda_{hh}(\infty) &= e^{-R_0^{hh} f_h} \end{aligned} \tag{3.61}$$

A randomly chosen type 1 human host will not be infected during a large outbreak only if it escapes getting infected from both the primary and secondary transmissions.

$$f_{h,1} = \nu \left(1 - e^{-\lambda f_a} e^{-R_0^{hh} f_h}\right) \tag{3.62}$$

The prefactor ν is to normalize the relative size of the epidemic by size of the population of type 1 human hosts. Similarly, we can calculate the size of the epidemic in type 2 hosts.

$$f_{h,2} = (1 - \nu) \left(1 - e^{-R_0^{hh} f_h}\right) \tag{3.63}$$

The total size of the epidemic in the human population is obtained by adding equations 3.62 and 3.63

$$\begin{aligned} f_h &= f_{h,1} + f_{h,2} \\ f_h &= 1 - \left(1 - \nu + \nu e^{-\lambda f_a}\right) e^{-R_0^{hh} f_h} \end{aligned} \tag{3.64}$$

The solution of the implicit equation 3.60 feeds in to equation 3.64 whose solution can then be used to solve equations 3.62 and 3.63. In the absence of secondary transmissions, i.e, $R_0^{hh} = 0$, the epidemic in the type 1 hosts would only consist of primary infections. Let this

fraction of infected hosts be denoted by $f^0_{h,p}$, which can be obtained by setting R^{hh}_0 to 0 in equation 3.62.

$$f^0_{h,p} = \nu \left(1 - e^{-\lambda f_a}\right) \tag{3.65}$$

Immediately comparing equations 3.62 and 3.65, we can assert that

$$f^0_{h,p} \leq f_{h,1} \tag{3.66}$$

with the equality holding for $R^{hh}_0 = 0$. Note that $f_{h,p} \neq f^0_{h,p}$ since $f^0_{h,p}$ is the size of the epidemic in the absence of human to human transmissions whereas $f_{h,p}$ is the size of the epidemic when both forces of infection are active. In the latter scenario, the two forces of infection would be competing for a susceptible. Thus, the proportion of the epidemic caused by primary infections would be reduced as compared to the case where only the primary transmission is active.

$$f_{h,p} \leq f^0_{h,p} \tag{3.67}$$

For the last part of the analysis, consider a randomly chosen infected type 1 human host i. This host is exposed to both primary and secondary forces of infection. Let $T^{(i)}_{h,p}$ be the time when this host receives disease via a primary transmission. Similarly, let $T^{(i)}_{h,s}$ be the time when the host receives disease via a secondary transmission. If $T^{(i)}_{h,p} < T^{(i)}_{h,s}$, a primary infection is realized else a secondary infection is realized. Note that the idea of multiple transmissions is a mathematical construct rather than a biological realism. A host that has already been infected and recovered can not be infected again (in the SIR framework). But the host is still subjected to the second force of infection which can result in another successful (albeit redundant) transmission. From the analogy with reaction kinetics [53], it is important to know which transmission reaction fired first since that would determine

whether the infection was primary or secondary. We can now write down an expression for the relative size of the epidemic consisting of primary infections.

$$
\begin{aligned}
f_{h,p} &= \nu \cdot \mathbb{P}[T^{(i)}_{h,p} < T^{(i)}_{h,s}] \\
&= \nu\Big(\mathbb{P}[T^{(i)}_{h,p} < T^{(i)}_{h,s}, T^{(i)}_{h,s} = \infty] + \mathbb{P}[T^{(i)}_{h,p} < T^{(i)}_{h,s}, T^{(i)}_{h,s} < \infty]\Big) \\
&= \nu\, e^{-R_0^{hh} f_h} \left(1 - e^{-\lambda f_a}\right) f_{h,1} + \nu\, \mathbb{P}[T^{(i)}_{h,p} < T^{(i)}_{h,s} < \infty] \\
&\geq f_{h,p\backslash s}
\end{aligned}
\tag{3.68}
$$

where

$$
f_{h,p\backslash s} \equiv \nu\, e^{-R_0^{hh} f_h} \left(1 - e^{-\lambda f_a}\right) f_{h,1} \tag{3.69}
$$

Combining equations 3.66, 3.67 and 3.69, we get

$$
f_{h,p\backslash s} \;\leq\; f_{h,p} \;\leq\; f^0_{h,p} \;\leq\; f_{h,1} \tag{3.70}
$$

where the equality holds for $R_0^{hh} = 0$.

3.5.3 Primary vs Secondary

Figure 3.13 shows the average number of primary and secondary infections occurring during a large outbreak for different values of ν and R_0^{hh}. The solutions were obtained by solving the deterministic equations (eq. 3.54). The curve for the primary infections will always be non-decreasing with ν. This follows from intuition that as more and more susceptible hosts become at risk, the number of primary infections will also increase. The fact that the effective R_0 for the A-H transmissions, i.e., R_0^{ah} also increases with ν compounds the effect. The secondary infections on the other hand exhibit non-monotonicity in some regions of parameter space. This can be attributed to the love-hate relationship between the two forces of infection (ah and hh) acting on susceptible human hosts. On one hand, the

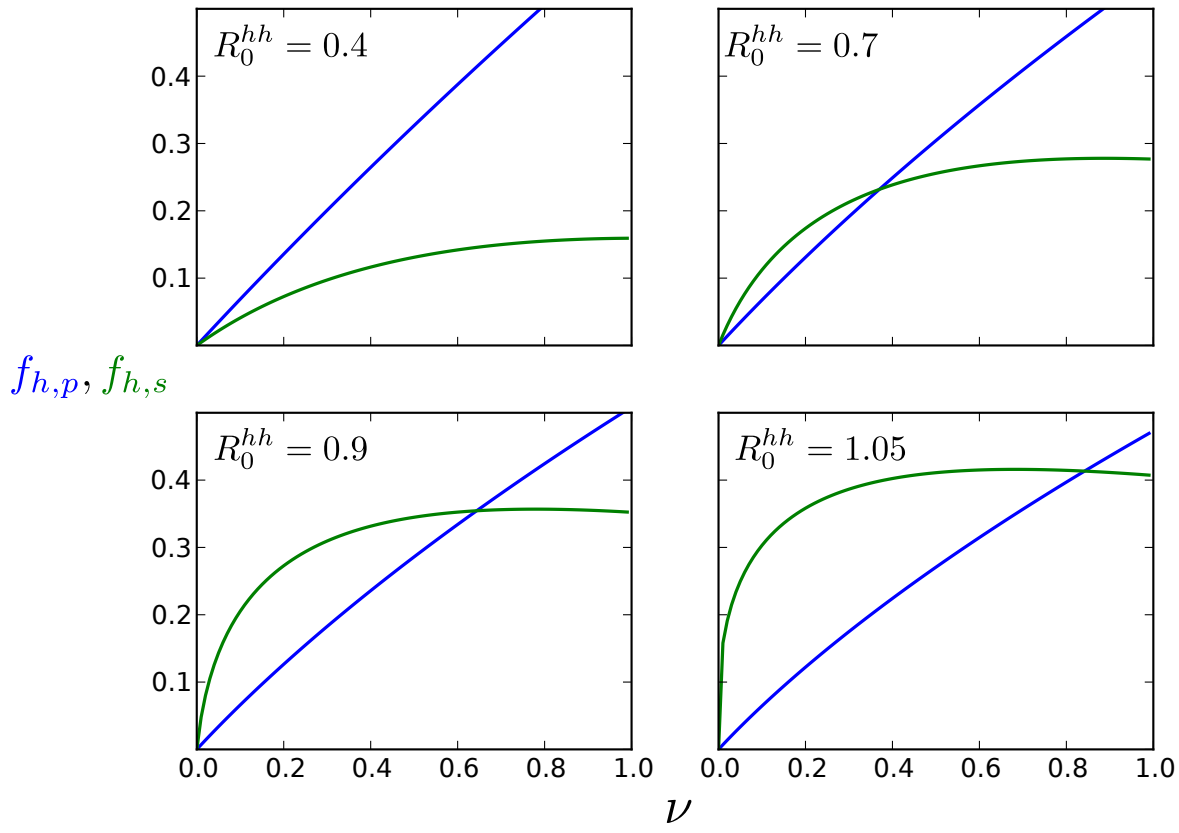


Figure 3.13: Fraction of human hosts infected during a large outbreak via a primary transmission ($f_{h,p}$, blue), and secondary transmission ($f_{h,s}$, green) plotted for different values of ν and R_0^{hh}. Analytical solution obtained from solving the deterministic system. Remaining parameters for the plots: $N_a = N_h = 1000, \beta_{aa} = 2.0, \hat{\beta}_{ah} = 1.5, \gamma_a = \gamma_h = 1.0$

secondary infections cannot occur unless there are primary infections. Thus, for small values of ν, there is a strong correlation between the number of primary and secondary infections. On the other hand, as ν increases, the two forces start competing for the same susceptible hosts. Depending on the model parameters, either of the two forces can dominate in different regions of the phase space which leads to the rich behavior for the number of secondary infections.

3.5.4 Bifurcation point

As evident from figure 3.13, the curves $f_{h,p}$ and $f_{h,s}$ when plotted against ν may or may not intersect apart from $\nu = 0$. Here we calculate the condition under which the bifurcation would occur creating a second point of intersection. Since we do not have an explicit expression for $f_{h,p}$ or $f_{h,s}$, the solution is not rigorous. But numerical experiments over a large parameter ranges have revealed that the solution does hold. The solution assumes that both $f_{h,p}$ and $f_{h,s}$ are concave functions of ν. For small values of R_0^{hh}, we can assert that the secondary infections would be smaller than primary infections for all values of ν. Thus $\nu = 0$ would be the only solution. As we increase R_0^{hh}, a bifurcation would occur at $\nu = 0$ and a second solution would emerge. At the bifurcation point, the slope of $f_{h,p}$ and $f_{h,s}$ would be equal. Thus, the bifurcation condition is

$$\left.\frac{\partial f_{h,p}}{\partial \nu}\right|_{\nu=0} = \left.\frac{\partial f_{h,s}}{\partial \nu}\right|_{\nu=0} \tag{3.71}$$

Since we don't have an analytical expression for $f_{h,p}$, we will work with equation 3.70. Assuming $R_0^{hh} < 1$, from equation 3.64 we get

$$\begin{aligned} f_h|_{\nu=0} &= 0 \\ \left.\frac{\partial f_h}{\partial \nu}\right|_{\nu=0} &= \frac{1}{1 - R_0^{hh}} \end{aligned} \tag{3.72}$$

Using the above solutions in 3.65 and 3.68, we obtain

$$\left.\frac{\partial f_{h,p\backslash s}}{\partial \nu}\right|_{\nu=0} = \left.\frac{\partial f^0_{h,p}}{\partial \nu}\right|_{\nu=0} = 1 \tag{3.73}$$

From equation 3.73 and 3.70, we obtain.

$$\left.\frac{\partial f_{h,p}}{\partial \nu}\right|_{\nu=0} = 1 \tag{3.74}$$

For $f_{h,s}$,

$$\left.\frac{\partial f_{h,s}}{\partial \nu}\right|_{\nu=0} = \left.\frac{\partial f_h}{\partial \nu}\right|_{\nu=0} - \left.\frac{\partial f_{h,p}}{\partial \nu}\right|_{\nu=0}$$
$$= \frac{R_0^{hh}}{1 - R_0^{hh}} \tag{3.75}$$

Equating 3.74 and 3.75, we obtain $R_0^{hh} = 1/2$ is the bifurcation point where the two slopes are equal. For $R_0^{hh} < 1/2$, the number of secondary transmissions will always be smaller than primary ones for $\nu > 0$. For $R_0^{hh} > 1/2$, the two curves will either intersect or $f_{h,s}$ will be strictly greater than $f_{h,p}$. We were unable to calculate analytically the point $\nu^\star$ of intersection of the two curves or the point in parameter space where the point of intersection disappears.

3.6 Discussion

We have analyzed a stochastic model of coupled infection dynamics in an animal-human metapopulation using the theory of multitype branching processes. Our results follow from well-established theory [11, 12, 15, 51, 52] that we applied to the problem of zoonoses. We have described spillover from animal to human populations, but such a model – or a variant of it, perhaps with a different form of interspecies coupling – would be applicable to other cross-species infections, such as among different animal hosts. The coupling of animal and human infectious disease dynamics results in important changes to the structure of outbreaks in human populations as compared to those in a human-only SIR model.

The statistical quantities that we have calculated in this work provide important insights into spillover and zoonoses. For instance, the probability of spillover (section 3.3.2) is strictly less than 1 because the outbreak can die out in the animal population before any primary human infections occur. The distribution of first passage times is useful for

understanding the relevant timescales of spillover and their stochastic fluctuations. This serves two important purposes: first, it indicates whether demography should be factored into the model (i.e., whether spillover will take place on a timescale fast compared to demographic changes), and more crucially, it suggests strategies for optimal surveillance in the field to pinpoint the relevant timescales and surveillance frequencies needed to identify emerging zoonotic infections. Lastly the distribution of disease prevalence in the animal population at the moment of spillover (section 3.3.3) highlights the intrinsic challenges to parameter estimation in order to build predictive models of cross-species spillover based on prevalence information.

The critical threshold and statistics of human outbreaks allow us to do a comparative analysis between the simple SIR and the coupled SIR presented in this work. Unlike the simple SIR, in our multi-species model the expected outbreak size diverges if either $\mathcal{R}_0^{aa}$ or $\mathcal{R}_0^{hh}$ exceeds 1 (section 3.4.2). Thus, large outbreaks in the human population are possible even if $\mathcal{R}_0^{hh} < 1$, emphasizing the potential importance of *spillover-driven large outbreaks*. This could have important ramifications for zoonotic diseases where human-to-human transmission is not the crucial determinant of the epidemic outcome such as rabies, Nipah, Hendra and Menangle [44, 49]. At the multicritical threshold, the outbreak sizes for the epidemics in the animal and human populations diverge simultaneously, resulting in a new universality class with a different scaling behavior. The animal-human coupling enhances the probability of longer chains ($P(n) \sim n^{-5/4}$ as compared to $P(n) \sim n^{-3/2}$), which could allow for greater opportunity for pathogen adaptation to human hosts [27, 56]. Furthermore, depending on where the system is in parameter space, the scaling of the *average* outbreak in the human population sizes can vary significantly (eqs. (3.43b) and (3.45)).

Our analysis suggests the need to be precise with terminology arising in the study of cross-species outbreaks. Various classification schemes previously proposed have delineated

among different zoonotic infections based on how infectious they are in the human population, e.g., stages II, III and IV discussed in [27, 43]. But we have shown that zoonotic infections in all three stages can support large outbreaks in the human population if driven sufficiently hard by an animal outbreak. In addition, the term 'stuttering chain' has been used in the literature [27, 56] to describe a chain of infections starting from a single infectious host that goes extinct without affecting a significant fraction of the host population. For the single-type SIR model, the term is synonymous with 'small outbreak' as we have defined here, and the epidemic threshold is the point in parameter space at which the average length of one such chain diverges. But in our multitype SIR model, the term 'stuttering chain' can not be used interchangeably with 'small outbreak'. Since multiple introductions can occur in the human population, an outbreak is small if and only if (1) a finite number of distinct infection chains occur in the human population, and (2) all such chains stutter to extinction. A large outbreak in the human hosts occurs when any one of these conditions is violated. Specifically, a spillover-driven large outbreak occurs when the number of infection chains diverges, which can happen if $\mathcal{R}_0^{aa} > 1$. Separately, the length of any one such chain can diverge if $\mathcal{R}_0^{hh} > 1$.

The community has advocated 'model-guided fieldwork' [27, 57, 58], as well as increased collaboration between public health scientists and ecologists in developing integrated approaches to predicting and preventing zoonotic epidemics[49]. Mathematical analysis needs to play a central role in such activities, in order to assess the implications of model assumptions. In this paper, we have endeavored to systematically characterize the behavior of a simple model system as a function of model parameters, and various extensions of this work are possible. Within the model itself, one could relax the assumption of homogeneous, full mixing within each population in order to investigate the role of heterogeneous mixing on complex contact networks [39, 40, 59]. Alternatively, one could develop models of

processes outside of the scope of the current model, to address factors such as the ecology of interactions between wildlife and domesticated animals, the encroachment of human development into animal habitats, the evolution of virulence, the propensity for pathogens to successfully jump across species, and the efficacy of various control strategies. We envision the parameters of our cross-species infection model to be the interface to that broader class of models, which would specify how cross-species infection parameters change over time as a function of ecological, evolutionary and immunological factors. In the current work, we have sought to identify key aspects of phenomenology, highlight the role of important processes, and suggest further inquiry into particular systems of interest, in order to help frame more complex and comprehensive descriptions of zoonotic infection.

CHAPTER 4

Chapter 4: Epidemics in metapopulation networks

Abstract [1]

Metapopulation models of infectious disease spread with explicit node dynamics have been a subject of research for many decades, although most research on stochastic formulations of these models has resided in the realm of simulation rather than analysis. In this chapter, we present a formalism for calculating the distribution of outbreak sizes, the epidemic threshold and relative final size of large outbreaks analytically in a weakly connected metapopulation network. This new formalism does not make any assumption about the network structure as has been the case in previous work. We demonstrate the applicability of our approach through a pedagogical example of one-dimensional lattice of coupled populations and a contrived metapopulation network for a livestock supply chain. Our analytical framework can aid in constructing 'risk maps' for disease outbreaks in farm or urban networks.

[1] This chapter and the material in Appendix E are being prepared for submission with co-authors David J. Schneider and Christopher R. Myers.

4.1 Introduction

The simplest models of infectious disease dynamics assume a single population of fully-mixed hosts, while metapopulation models relax that central assumption, by describing a set of fully-mixed subpopulations that are coupled together in some prescribed fashion, as illustrated schematically in figure 4.1. Metapopulation models are commonly used to describe disease spread (and other ecological processes) in populations with clusters of spatially distributed hosts. This includes models of a wide range of infectious diseases that pose a threat to public health (e.g., H1N1 [60], SARS [61], influenza [62]) or disrupt agricultural economic activity (e.g., foot and mouth disease [21, 63, 64]). The spread of disease among subpopulations is typically a consequence of host movement [65, 66] or pathogen dispersal [67, 68], but simplifying assumptions regarding inter-population spread are often made, such as subsuming the details of the contact process in a simple coupling coefficient [65, 69–72] or a transmission kernel [21, 73, 74]. This simplification allows one to analyze disease spread, typically either deterministically using a system of coupled differential equations or stochastically via a coupled birth-death process [12]. Metapopulations describe populations with two levels of structure: at the level of individual hosts, and at the level of the *metapopulation network*, which relates how different subpopulations are coupled to each other. Since interactions in infectious disease are asymmetric (infectious hosts infect susceptible hosts, but not *vice versa*), metapopulation networks are directed graphs.

Prior analytical work in characterizing stochastic disease dynamics in metapopulations has focused primarily on the cases of fully connected and strongly connected metapopulation networks [10, 18, 75]. In the former, each subpopulation is coupled to all others; in the latter, each subpopulation can be reached from all others by following directed edges. An

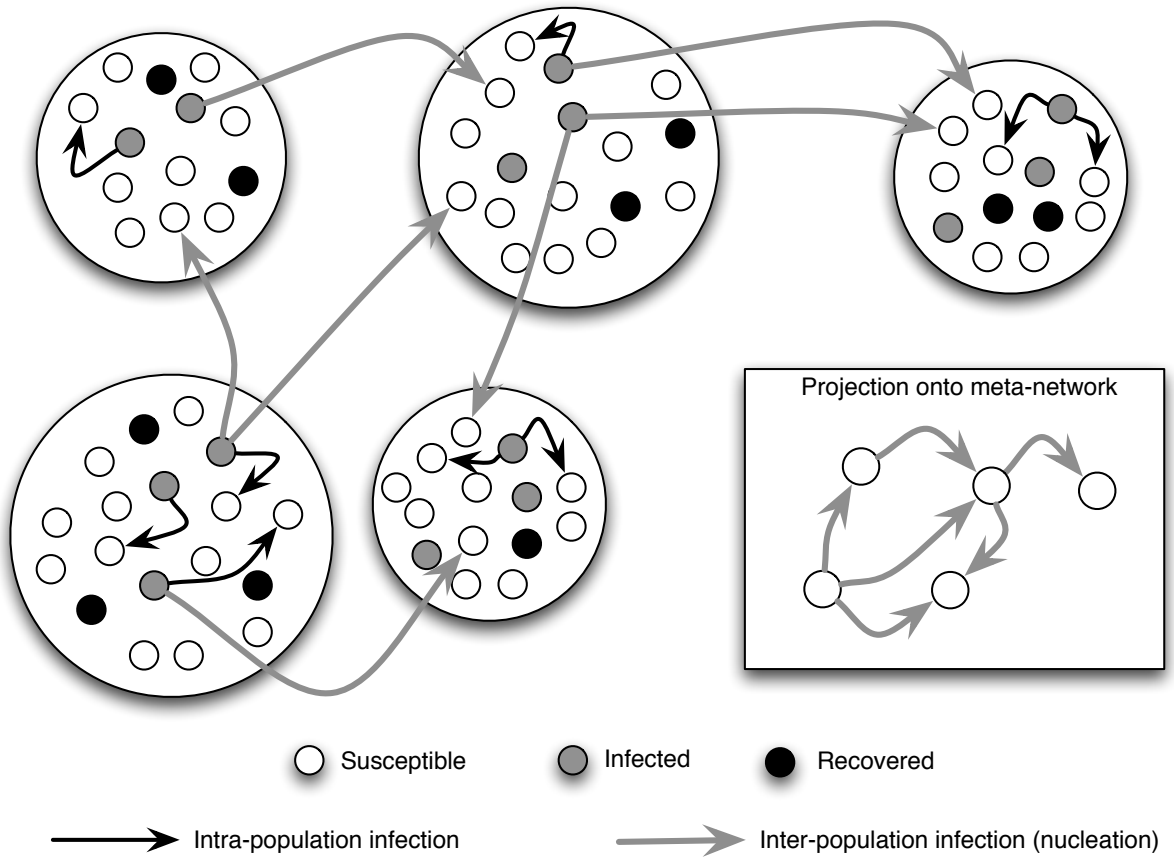


Figure 4.1: Schematic depiction of the metapopulation networks analyzed here. Each node i in the metapopulation network includes a well-mixed population of N_i hosts, with infection described by an SIR model relating populations of Susceptible (S), Infectious (I), and Recovered (R) hosts. Infectious hosts are able to infect susceptible hosts either within a given subpopulation (black arrows) or between subpopulations (gray arrows). Disease spread is defined only in terms of individual hosts, but some quantities are usefully computed by projecting onto the meta-network connecting subpopulations (inset), where we can ask, for example, whether any host within a given subpopulation becomes infected.

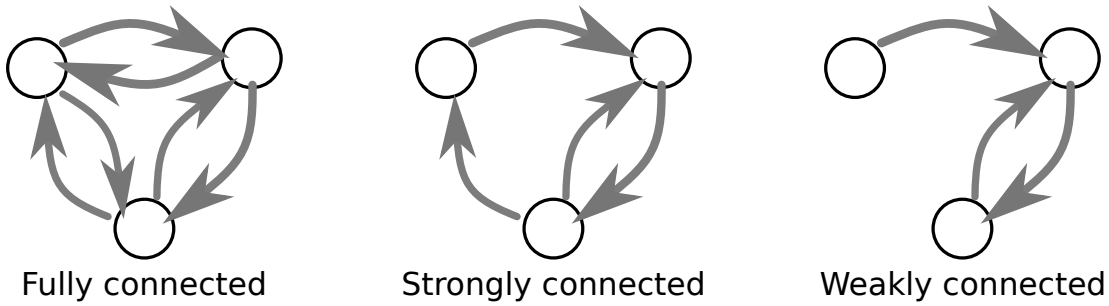


Figure 4.2: Schematic depiction of the three types of metapopulation networks based on node connectivity.

important subclass of metapopulation networks that has not received much attention, which we focus on here, is the case of strictly weakly connected metapopulation networks: directed networks that are weakly connected but not strongly connected. We will henceforth use the term 'weakly connected' to imply 'strictly weakly connected', even though a strongly connected graph is also weakly connected but not strictly so. Figure 4.2 depicts different types of directed networks. An alternative terminology has been used in the infectious disease literature, distinguishing between *reducible* contact patterns (strictly weakly connected) and *irreducible* ones (strongly connected) [76]. Weakly connected networks can be decomposed into a network of strongly connected components; see figure 4.3. Disease dynamics on weakly connected networks can exhibit interesting behavior where local mixing in the strongly connected components can act as a breeding reservoir for the pathogen and lead to disease persistence. The likelihood of disease extinction in such networks depends on the structure of the network and the node where the outbreak was initiated. These networks also impose unidirectionality in disease transmission, a feature that is also observed in some real life systems. Cholera transmission along dendritic river systems [77, 78], foot and mouth transmission in farm networks [21, 63], computer virus transmission on the internet subnetworks [79, 80] are some of the epidemiological examples that constitute this class of problems.

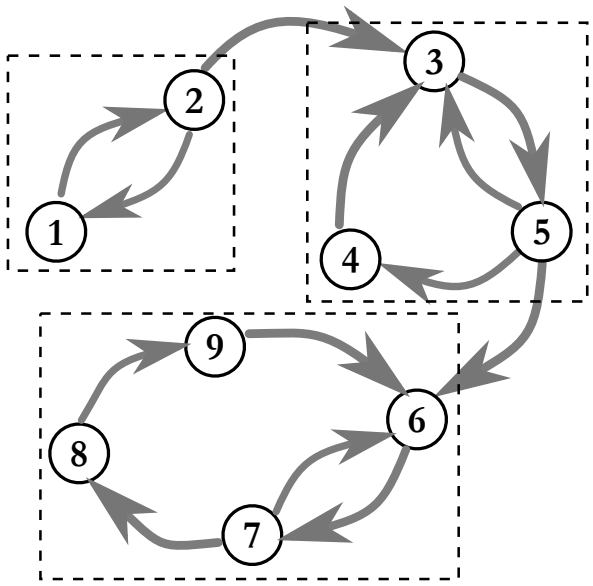


Figure 4.3: Decomposition of a weakly connected graph in to strongly connected components.

We define a risk map on a metapopulation network to be a quantitative attribution of epidemic risk to each node in the network. A set of risk maps can be defined as follows:

1 The probability of disease invasion in any given node of the network given its point of origin in the network.

2 The probability of a large outbreak in any given node of the network given its point of origin in the network.

3 The probability of a node being the point of origin given that infection was detected in some part of the network.

In a weakly connected metapopulation network, multiple independent pathways potentially exist for the disease to invade a given part of the network. For instance, the supply chain of beef consists of farms, cattle markets, feedlots and slaughterhouses as a weakly connected network. An instance of infection in the slaughterhouse that is untraceable can arrive from any of the large number of sources in the supply chain. For

such a system, the notion of risk maps is important from the perspective of detection and control. Through risk maps, the observed information (of disease being detected at a particular node) can be combined with network topology and the model of disease spread to provide a quantitative picture of impending epidemics.

In this work, we employ the analytical framework of epidemic percolation networks [40] and introduce a formalism to calculate the distribution of outbreak sizes, the epidemic threshold and relative size of large outbreaks in an arbitrary weakly connected network. Our formalism builds in part on that previously developed to describe Reed-Frost processes (a discrete-time stochastic model of disease spread) on reducible contact networks [76]. We are concerned here with SIR processes on such networks. While the final size of large outbreaks for equivalent Reed-Frost and SIR models are identical, the distribution of sizes of small outbreaks, the probability of spillover and that of a large outbreak are different, and to our knowledge, solution for these quantities in an SIR model on a weakly connected (reducible) network has not been previously published.

4.2 Model Description

We introduce a coupled SIR model with M population nodes where for every pair of nodes i and j that have a directed edge from i to j, the host population is homogeneously mixed, i.e., every infectious host in node i has the same probability per unit time of infecting any susceptible host in node j. Note that i and j can be the same index in which case the notion of being fully mixed is trivial. Let this rate of infectious contact be $\rho_{ij}\beta_i/N_j$ where N_j is the host population in node j, ρ_{ij} is the coupling coefficient ($\rho_{ii} = 1$) and β_i is the cumulative rate of infectious contact for hosts in node i. We assume that β_i is a random variable drawn from a known distribution $F_1(\beta_i)$. Similarly, we define τ_i as the

infectious period for hosts in node i where τ_i is also a random variable drawn from a known distribution $F_2(\tau_i)$. For every node i, we shall define $\mathcal{R}_0^{(i)} = \mathbb{E}[\beta_i]\,\mathbb{E}[\mu_i]$ as the Basic Reproduction Number local to that particular node. This quantity represents the average number of new infections produced within node by an infectious host when introduced in a fully susceptible host population.

For a given node, an outbreak occurs when one or more individuals in the node are infected via external importation of infection [40]. The size of an outbreak refers to the total number of individuals infected at the end of the infection process and its relative size is the size divided by the total host population. The outbreak is self-limited if it has zero relative size in the limit of infinite system size. On the other hand, an outbreak with strictly positive relative size is defined as large. In this analysis, we are interested in calculating the distribution of outbreak sizes, the probability of having a large outbreak and its relative size in an arbitrary metapopulation network.

The Epidemic Percolation Network (EPN) introduced by Kenah and Robins [40, 55] is a convenient tool for analyzing the SIR process asymptotically at long times ($t \to \infty$). Although this procedure does not provide any insight in to the dynamics of the process, it illuminates the impact of the outbreak, i.e., whether the disease dies out quickly or leads to a large outbreak. This analysis is valid in the limit of infinite system size. For the most part, we will borrow the notation described in [40] but modified to accommodate our metapopulation system. We will refer to the population in node i as $\mathcal{P}_i$. In the original epidemic percolation network, nodes refer to individual hosts, which we refer to here as "host nodes", so as to avoid any confusion with the more coarse-grained nodes in the metapopulation network that we have been discussing. (In addition, we denote as a "host edge" any connection between host nodes.) The underlying contact network in our system is a complete graph with $N = \sum_{k=1}^{M} N_k$ host nodes such that N_i randomly chosen host

nodes belong to $\mathcal{P}_i$. The complete graph assumption comes from the fact that the host population is fully mixed. The percolation network on the other hand is a semi-directed graph where each host node is connected to the network via incoming and outgoing directed edges. The percolation procedure when performed on an arbitrary network also leads to undirected edges, but they do not appear in the case of a fully mixed network [55]. The directed graph represents the transmission pathways followed by the disease while moving from host to host. The details of the forthcoming calculations pertaining to the percolation network are shown in the Appendix E.1.

Let $P(m_1,m_2,\ldots,m_M;n_1,n_2,\ldots,n_M)$ be the probability that a randomly chosen host node in $\mathcal{P}_i$ has $m_1, m_2, \ldots, m_M$ incoming edges and $n_1, n_2, \ldots, n_M$ outgoing edges to other host nodes belonging to $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_M$ in the percolation network. Let $G_i(\boldsymbol{x}, \boldsymbol{y})$ be the PGF for the degree distribution of incoming edges and outgoing edges, i.e.,

$$G_i(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{m}=0}^{\infty}\sum_{\boldsymbol{n}=0}^{\infty} P(m_1,m_2,\ldots,m_M;n_1,n_2,\ldots,n_M) x_1^{m_1} x_2^{m_2}\ldots x_M^{m_M}\, y_1^{n_1} y_2^{n_2}\ldots y_M^{n_M} \quad (4.1)$$

Using the framework of epidemic percolation networks (appendix E.1), the PGF is given by

$$G_i(\boldsymbol{x}, \boldsymbol{y}) = \exp\left[\sum_{j=1}^{M} \rho_{ji}\mathcal{R}_0^{(j)}\eta_j(x_j-1)/\eta_i\right] \cdot \int_0^{\infty} \exp\left[\sum_{j=1}^{M} \rho_{ij}\mathbb{E}[\beta_i](y_j-1)\tau_i\right] dF(\tau_i) \quad (4.2)$$

where $\eta_i = \lim_{N\to\infty} N_i/N$. In EPN, the distribution of in-degree and out-degree of a node are independent. We define $G_i^{in}(\boldsymbol{x})$ and $G_i^{out}(\boldsymbol{y})$ as the PGFs for the indegree and outdegree distributions contained in $G_i(\boldsymbol{x}, \boldsymbol{y})$. Then

$$G_i(\boldsymbol{x}, \boldsymbol{y}) = G_i^{in}(\boldsymbol{x}) G_i^{out}(\boldsymbol{y}) \quad (4.3)$$

where

$$\begin{aligned} G_i^{in}(\boldsymbol{x}) &= \exp\left[\sum_{j=1}^{M} \rho_{ji}\mathcal{R}_0^{(j)}\eta_j(x_j-1)/\eta_i\right] \\ G_i^{out}(\boldsymbol{y}) &= \int_0^{\infty} \exp\left[\sum_{j=1}^{M} \rho_{ij}\mathbb{E}[\beta_i](y_j-1)\tau_i\right] dF(\tau_i) \end{aligned} \quad (4.4)$$

Let $z_{i\to j}$ be the average number of outgoing edges of a host in $\mathcal{P}_i$ that lead to a random host in $\mathcal{P}_j$. Similarly, let $z_{i\leftarrow j}$ be the average number of incoming edges from a random host in $\mathcal{P}_j$:

$$\begin{aligned} z_{i\to j} &= \left.\frac{\partial G_i(\boldsymbol{x},\boldsymbol{y})}{\partial y_j}\right|_{(\mathbf{1},\mathbf{1})} = \rho_{ij}\mathcal{R}_0 \\ z_{i\leftarrow j} &= \left.\frac{\partial G_i(\boldsymbol{x},\boldsymbol{y})}{\partial x_j}\right|_{(\mathbf{1},\mathbf{1})} = \rho_{ji}\mathcal{R}_0\frac{\eta_j}{\eta_i} \end{aligned} \tag{4.5}$$

Let $G_{i\to j}(\boldsymbol{x},\boldsymbol{y})$ be the PGF for the degree distribution of a host node in $\mathcal{P}_j$ reached by going forward along a randomly chosen directed host edge coming from a host node in $\mathcal{P}_i$, excluding the host edge used to reach the host node. Then $G_{i\to j}(\boldsymbol{x},\boldsymbol{y})$ can be calculated as

$$\begin{aligned} G_{i\to j}(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{z_{j\leftarrow i}}\frac{\partial G_j(\boldsymbol{x},\boldsymbol{y})}{\partial x_i} \\ &= \frac{1}{z_{j\leftarrow i}}\frac{\partial G_j^{in}(\boldsymbol{x})}{\partial x_i}G_j^{out}(\boldsymbol{y}) \end{aligned} \tag{4.6}$$

Similarly, we can define $G_{i\leftarrow j}(\boldsymbol{x},\boldsymbol{y})$ as the PGF for the degree distribution of a host node in $\mathcal{P}_j$ that is reached by following a randomly chosen directed host edge in reverse from a host node in $\mathcal{P}_i$:

$$\begin{aligned} G_{i\leftarrow j}(\boldsymbol{x},\boldsymbol{y}) &= \frac{1}{z_{j\to i}}\frac{\partial G_j(\boldsymbol{x},\boldsymbol{y})}{\partial y_i} \\ &= \frac{1}{z_{j\to i}}G_j^{in}(\boldsymbol{x})\frac{\partial G_j^{out}(\boldsymbol{y})}{\partial y_i} \end{aligned} \tag{4.7}$$

Note the following relations:

$$\begin{aligned} G_i(\mathbf{1},\boldsymbol{y}) &= G_{i\to j}(\mathbf{1},\boldsymbol{y}) = G_j^{out}(\boldsymbol{y}) \\ G_i(\boldsymbol{x},\mathbf{1}) &= G_{i\leftarrow j}(\boldsymbol{x},\mathbf{1}) = G_j^{in}(\boldsymbol{x}) \end{aligned} \tag{4.8}$$

Until this point, we have not introduced anything new except for applying the framework of epidemic percolation networks to the metapopulation case and deriving some well

known results. In the next section, we present the formalism that is used for strongly connected networks and calculate the the distribution of outbreak sizes, probability of a large outbreak, the epidemic threshold and the relative size of large outbreaks.

4.3 Epidemics on strongly connected metapopulation networks

4.3.1 Distribution of outbreak sizes

It has been established [40] that the distribution of outbreak sizes initiating at a randomly chosen host p is the same as the distribution of the out-components connected to that host node in the percolation network. The out-component of the host p is the set of all hosts (including p) that can be reached by following outgoing or undirected edges in the host network. We define $H_i^{out}(\boldsymbol{y})$ as the PGF for the distribution of the out-component sizes of a random host in $\mathcal{P}_i$. This distribution can be calculated as follows [75, 81]

$$\begin{aligned} H_i^{out}(\boldsymbol{y}) &= y_i G_i(\mathbf{1}, \tilde{\boldsymbol{H}}^{\boldsymbol{out}}(\boldsymbol{y})) \\ \tilde{H}_j^{out}(\boldsymbol{y}) &= y_j G_{i\to j}(\mathbf{1}, \tilde{\boldsymbol{H}}^{\boldsymbol{out}}(\boldsymbol{y})) \end{aligned} \tag{4.9}$$

where $\tilde{\boldsymbol{H}}^{\boldsymbol{out}}(\boldsymbol{y}) = (\tilde{H}_1^{out}(\boldsymbol{y}), \tilde{H}_2^{out}(\boldsymbol{y}), \cdots, \tilde{H}_M^{out}(\boldsymbol{y}))$. Using (4.8), we can establish the following

$$\begin{aligned} H_i^{out}(\boldsymbol{y}) &= \tilde{H}_i^{out}(\boldsymbol{y}) \\ H_i^{out}(\boldsymbol{y}) &= y_i G_i^{out}(\boldsymbol{H}^{\boldsymbol{out}}(\boldsymbol{y})) \end{aligned} \tag{4.10}$$

4.3.2 Epidemic Threshold

The epidemic threshold is the point at which a giant component emerges in the network whose size scales with the size of the network. It is also the point at which the expected outbreak size diverges. Below the epidemic threshold, let s_{ij} be the expected size of outbreak in $\mathcal{P}_j$ given that the outbreak started in node i.

$$\begin{aligned} s_{ij} &= \left.\frac{\partial H_i^{out}(\boldsymbol{y})}{\partial y_j}\right|_{\boldsymbol{y}=\mathbf{1}} \\ &= \delta_{ij} + \sum_{k=1}^{M} \rho_{ij} R_0^{(i)} s_{kj} \end{aligned} \tag{4.11}$$

Let $\boldsymbol{s}$ represent the $M \times M$ matrix $[s_{ij}]$. Then $\boldsymbol{s}$ satisfies the following equation

$$\boldsymbol{s} = \mathbb{I} + \mathbf{R} \cdot \boldsymbol{\rho} \cdot \boldsymbol{s} \tag{4.12}$$

where $\mathbb{I}$ is an identity matrix, $\mathbf{R}$ is a diagonal matrix with elements $(R_{01}, R_{02}, \ldots, R_0^{(M)})$ and $\boldsymbol{\rho}$ is the matrix $[\rho_{ij}]$. Let $\mathbf{A} = \mathbf{R} \cdot \boldsymbol{\rho}$. The equation (4.12) can be solved as

$$\boldsymbol{s} = (\mathbb{I} - \mathbf{A})^{-1} \tag{4.13}$$

The inverse of the matrix is well defined if $\det(\mathbb{I} - \mathbf{A}) \neq 0$. The case of $\det(\mathbb{I} - \mathbf{A}) = 0$ marks the onset of an epidemic when a giant component emerges in the network that scales with the system size [39, 40]. $\mathbf{A}$ is the Next Generation Matrix (NGM) of the system [82] and its largest eigenvalue $\lambda_{\max}$ sets the epidemic threshold, i.e., $\lambda_{\max} = 1$. A similar calculation exists in [75, 81] for percolation on multitype networks.

4.3.3 Probability of a large outbreak

When the network epidemic threshold is breached ($\lambda_{\max} > 1$), there exists a unique giant component in the network that includes all the hosts that are infected in event of a large

outbreak [39, 55]. The probability of a large outbreak is the same as the probability that a randomly chosen host has a giant out-component in the percolation network [40, 55]. In the presence of the giant component the PGF $H_i^{out}(\boldsymbol{y})$ represents the distribution of the sizes of the small component connected to a randomly chosen host in $\mathcal{P}_i$, i.e., the distribution of sizes of a self-limited outbreak [39, 40] that initiated in $\mathcal{P}_i$. Thus, the probability of extinction π_i is the same as the probability of being in the small-component, i.e., $H_i^{out}(\mathbf{1})$. Thus

$$\pi_i = G_i^{out}(\boldsymbol{\pi}) \tag{4.14}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_M)$. This result can also be interpreted as the probability of extinction of a multitype Galton-Watson branching process (appendix B) with the offspring PGF $H_i^{out}(\boldsymbol{y})$ [83]. π_i represents the probability that an outbreak that starts in node i dies out. The probability of a large outbreak occurring in the network given that it starts in node i is simply one minus the probability of extinction, i.e, $1 - \pi_i$.

4.3.4 Relative size of a large outbreak

In the epidemic percolation network, the relative size of a large outbreak is the same as the probability that a randomly chosen host in the population has a giant in-component [40, 55]. The in-component of a host p is the set of all hosts (including p) that can be reached from p by following incoming or undirected edges in the host network. Let $H_i^{in}(\boldsymbol{x})$ be the PGFs for the distribution of in-components of a randomly chosen host in $\mathcal{P}_i$. As described in [40, 55], the distribution of connected in-components at a host node is generated by

$$H_i^{in}(\boldsymbol{x}) = x_i G_i^{in}(\boldsymbol{H^{in}}(\boldsymbol{x})) \tag{4.15}$$

In the epidemic percolation network, the size of the epidemic is equal to the probability that a randomly chosen host has a giant in-component [40, 55]. Let χ_i be the fraction of the host population in node i that is infected in the case of an epidemic. χ_i can be calculated as

$$\begin{aligned}\chi_i &= 1 - H_i^{in}(\mathbf{1}) \\ &= 1 - G_i^{in}(\mathbf{1} - \boldsymbol{\chi})\end{aligned} \tag{4.16}$$

where $\mathbf{1}$ is a vector of ones and $\boldsymbol{\chi} = (\chi_1, \chi_2, \ldots, \chi_M)$.

4.4 Epidemics on weakly connected metapopulation networks

The formalism described in the previous section (and proliferating in the literature [10, 18, 75, 81]) has a limitation that it is not applicable to a strictly weakly connected network. To demonstrate this limitation, we take the example of a simple metapopulation network with two nodes and one sided coupling (see figure 4.4). To simplify things, we choose

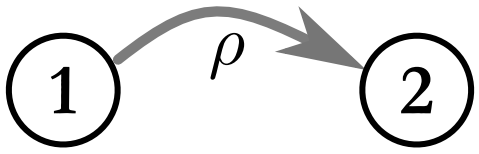


Figure 4.4: A simple two node system with one directional coupling

constant β_i for each node and exponentially distributed recovery times $F_2(\tau_i) \sim \text{Exp}(1/\mathbb{E}[\tau_i])$, making our model the *general epidemic* [26] or the Kermack-Mckendric SIR model on a metapopulation network. Let N_1, N_2 be the population of the two nodes. The basic reproduction number for the nodes are set to R_{01} and R_{02}. The calculations are performed in the limit of $N_1, N_2 \to \infty$ such that $N_1/(N_1 + N_2) \to \eta_1$ and $N_2/(N_1 + N_2) \to \eta_2$.

Using eq. 4.2, the probability generating functions for the two-node system simplify as follows

$$
\begin{aligned}
G_1(\boldsymbol{x}, \boldsymbol{y}) &= \frac{\exp\left[R_{01}(x_1 - 1)\right]}{1 - R_{01}\left[(y_1 - 1) + \rho(y_2 - 1)\right]} \\
G_2(\boldsymbol{x}, \boldsymbol{y}) &= \frac{\exp\left[\rho R_{01}(x_1 - 1)\,\eta_1/\eta_2 + R_{02}(x_2 - 1)\right]}{1 - R_{02}(y_2 - 1)}
\end{aligned}
\tag{4.17}
$$

The generating function $G_1(\boldsymbol{x}, \boldsymbol{y})$ can be split into $G_1^{in}(\boldsymbol{x})$ and $G_1^{out}(\boldsymbol{y})$, and similarly for $G_2(\boldsymbol{x}, \boldsymbol{y})$.

$$
\begin{aligned}
G_1^{out}(\boldsymbol{y}) &= \frac{1}{1 - R_{01}\left[(y_1 - 1) + \rho(y_2 - 1)\right]} \\
G_1^{in}(\boldsymbol{x}) &= \exp\left[R_{01}(x_1 - 1)\right] \\
G_2^{out}(\boldsymbol{y}) &= \frac{1}{1 - R_{02}(y_2 - 1)} \\
G_2^{in}(\boldsymbol{x}) &= \exp\left[\rho R_{01}(x_1 - 1)\,\eta_1/\eta_2 + R_{02}(x_2 - 1)\right]
\end{aligned}
\tag{4.18}
$$

The distribution of outbreak sizes in the two-node system is given by the following:

$$
\begin{aligned}
H_1^{out}(y_1, y_2) &= y_1 G_1^{out}(H_1^{out}(y_1, y_2), H_2^{out}(y_2)) \\
H_2^{out}(y_2) &= y_2 G_2^{out}(H_2^{out}(y_2))
\end{aligned}
\tag{4.19}
$$

For the probability of extinction π_1 and π_2, we have

$$
\pi_1 = H_1(1, 1), \quad \pi_2 = H_2(1) \tag{4.20}
$$

which can be obtained as follows:

$$
\begin{aligned}
\pi_1 &= \frac{1}{1 - R_{01}\left[(\pi_1 - 1) + \rho(\pi_2 - 1)\right]} \\
\pi_2 &= \frac{1}{1 - R_{02}(\pi_2 - 1)}
\end{aligned}
\tag{4.21}
$$

In the current case, let us assume that $R_{01}, R_{02} > 1$. π_i represents the probability that the outbreak goes extinct in all nodes accessible from node i. A node j is accessible from node i if there exists a directed path in the metapopulation network from i to j. Also, a given node is accessible from itself. While this is useful information, it is incomplete. In a given metapopulation network, one might be interested in the probability of extinction (or alternatively of a large outbreak) for a subgraph. The probability of extinction as calculated in the current framework only provides a lower-bound for the probability of extinction in the system. If there exist multiple outgoing paths from a given node, each of those paths will have a different probability of extinction. One of the key ideas that we present in this paper is a systematic way to calculate those probabilities of extinction for all possible paths in the network.

Next we calculate the epidemic threshold for this system using the methods outlined in section 4.3.2. The Next Generation Matrix is given by

$$\mathbf{A} = \begin{bmatrix} R_{01} & 0 \\ 0 & R_{02} \end{bmatrix} \cdot \begin{bmatrix} 1 & \rho \\ 0 & 1 \end{bmatrix} \tag{4.22}$$

The eigenvalues of this system are $[R_{01}, R_{02}]$ and the epidemic threshold is given by

$$\max\left(R_{01}, R_{02}\right) = 1 \tag{4.23}$$

The fallacy with this result is that it does not take into account the initial conditions, i.e, the node where the outbreak starts. The threshold given by eq. 4.23 is – incorrectly – a function of both R_{01} and R_{02} even though infection starting in node 2 is only governed by R_{02}. Further discrepancy is observed in the relative size of large outbreaks although that is not discussed here in detail.

Thus, the existing formalism falls short when applied to weakly connected networks. To extend the applicability, we adapt the notion of *reducible epidemics* that was introduced in

[76] for the multitype Reed-Frost processes. We introduce a formalism in the next section that describes SIR-type epidemic processes on weakly connected metapopulation networks. This formalism corrects the inconsistencies described in the preceding text and is applicable to an arbitrary weakly connected network, to account for the fact that all nodes may not be accessible from any given node.

4.4.1 Distribution of outbreak sizes

We define $H_{i\to j}(y)$ as the PGF for the distribution of the outbreak sizes in node j when the outbreak starts in node i:

$$H_{i\to j}(y) = \begin{cases} 1 & \text{if } i \neq j \text{ and } i \nrightarrow j, \\ G_i^{out}(\boldsymbol{H}_{\to \boldsymbol{j}}(y)) & \text{if } i \neq j \text{ and } i \to j, \\ yG_i^{out}(\boldsymbol{H}_{\to \boldsymbol{j}}(y)) & \text{if } i = j \end{cases} \tag{4.24}$$

where $i \to j$ implies that there exists a path from i to j and $i \nrightarrow j$ implies the negation. The symbol $\boldsymbol{H}_{\to \boldsymbol{j}}(y)$ is a shorthand for the vector $(H_{1\to j}(y), H_{2\to j}(y), \ldots, H_{M\to j}(y))$. Note that in the framework for strongly connected networks presented in section 4.3, the generating functions were defined over vector valued variables whereas in the present case, the generating functions are defined on a single variable. This is because the distribution in the current case only pertains to the outbreak size in node j whereas previously the generating functions calculated a joint distribution of outbreak sizes in all nodes in the network. Marginalizing the joint distribution naively is incorrect, as was demonstrated earlier (see eq. 4.21).

4.4.2 Epidemic Threshold

The decomposition of an arbitrary weakly connected graph into strongly connected components (SCC) provides some intuition for calculating the epidemic threshold. For an arbitrary M node graph, let this decomposition be $\mathcal{W} = \mathcal{S}_1 \cup \mathcal{S}_2 \ldots \cup \mathcal{S}_Q$. For any strongly connected subgraph, the epidemic threshold is the same as that derived in section 4.3.2.

For a weakly connected network $\mathcal{W}$, let λ_k be the largest eigenvalue of the NGM associted with the strongly-connected component $\mathcal{S}_k$. Let $\text{PATH}(\mathcal{S}_i, \mathcal{S}_j)$ be the set of all possible strongly connected subgraphs that can be visited in going from any node in $\mathcal{S}_i$ to an arbitrary node in $\mathcal{S}_j$. Note that we are not only interested in the shortest path, but *all possible* paths connecting nodes in $\mathcal{S}_i$ to nodes in $\mathcal{S}_j$.

$$\text{PATH}(\mathcal{S}_i, \mathcal{S}_j) = \{\mathcal{S}_k \mid \mathcal{S}_i \to \mathcal{S}_k \text{ and } \mathcal{S}_k \to \mathcal{S}_j\} \tag{4.25}$$

We also define $\text{PATH}(l, \mathcal{S}_j)$ as the set of all possible strongly connected subgraphs that can be visited in going from node l to any node in $\mathcal{S}_j$. Lastly, we define $\text{PATH}(l, m)$ as the set of all possible strongly connected subgraphs that can be visited in going from node l to node m.

To illustrate the definitions, note that in figure 4.3, we have the following

$$M = 8, Q = 3$$

$$\mathcal{S}_1 = \{1, 2\}, \mathcal{S}_2 = \{3, 4, 5\}, \mathcal{S}_3 = \{6, 7, 8, 9\}$$

For instance, consider all possible paths $\mathcal{S}_1 \to \mathcal{S}_3$. Since these would include all nodes in $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$,

$$\text{PATH}(\mathcal{S}_1, \mathcal{S}_3) = \{\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3\}$$

Let $\lambda_{i\to j}$ be defined as follows

$$\lambda_{i\to j} = \begin{cases} 0 & \text{if } i \nrightarrow j, \\ \max\left(\{\lambda_k \mid \mathcal{S}_k \in \text{PATH}(i,j)\}\right) & \text{otherwise} \end{cases} \tag{4.26}$$

Then, the epidemic threshold for a large outbreak in node j given that the infection starts in node i is given by

$$\lambda_{i\to j} = 1$$

If $\lambda_{i\to j} < 1$, it implies that the infection dynamics within any subgraphs that belong to the set $\text{PATH}(i,j)$ is not strong enough to cause a large outbreak in node j. A large outbreak in j may occur if a different eigenvalue (say $\lambda_{l\to j}$) is greater than 1. That implies that there is at least one subgraph belonging to the set $\text{PATH}(l,j)$ where the infection dynamics is supercritical.

4.4.3 Probability of spillover and large outbreak

As in the case of zoonoses (section 3.3.2), spillover in a node is defined as the event when one or more individuals are infected in the given node. Let $P_{i\to j}$ represent the probability that the disease spills into node j given that infection enters the system in node i:

$$P_{i\to j} = 1 - H_{i\to j}(0) \tag{4.27}$$

Similar to the probability of spillover, we define the probability of there being a large outbreak in node j given that infection starts in node i, as $Q_{i\to j}$. In the limit of large system size, the disease process in a strongly connected metapopulation network can either go extinct everywhere or lead to a large outbreak in every node in the network. Thus all $Q_{i\to j}$, $\forall j \in \mathcal{S}_k$ will be equal. In the EPN, the probability of an outbreak in $\mathcal{P}_j$ turning into

a large outbreak is the probability that the out-component of a randomly chosen host in $\mathcal{P}_i$ is infinite.

$$Q_{i\to j} = 1 - H_{i\to j}(1) \tag{4.28}$$

We thus have, $P_{i\to j} \geq Q_{i\to j}$. The equality holds only for the case where $H_{i\to j}(y) = 1$, i.e., $i \nrightarrow \mathcal{S}_k$. We also define the probability of extinction as

$$\pi_{i\to j} = 1 - Q_{i\to j} = H_{i\to j}(1) \quad \forall\, j \in \mathcal{S}_k \tag{4.29}$$

Note that the probability of extinction *includes* the probability that the outbreak never reaches node j. Lastly, the probability of there being a small outbreak in node j given that it starts in node i is calculated as

$$U_{i\to j} = P_{i\to j} - Q_{i\to j} \tag{4.30}$$

4.4.4 Relative size of the large outbreak

Similar to 4.4.1, let $H_{i\leftarrow j}(x)$ be the PGFs for the distribution of in-component sizes of a randomly chosen host in $\mathcal{P}_i$ that lie in $\mathcal{P}_j$:

$$H_{i\leftarrow j}(x) = \begin{cases} 1 & \text{if } i \neq j \text{ and } j \nrightarrow i, \\ G_i^{in}(\boldsymbol{H}_{\leftarrow \boldsymbol{j}}(x)) & \text{if } i \neq j \text{ and } j \to i, \\ xG_i^{in}(\boldsymbol{H}_{\leftarrow \boldsymbol{j}}(x)) & \text{if } i = j. \end{cases} \tag{4.31}$$

where $\boldsymbol{H}_{\leftarrow \boldsymbol{j}}(y)$ is a notation for the vector $(H_{1\leftarrow j}(y), H_{2\leftarrow j}(y), \ldots, H_{M\leftarrow j}(y))$.

Once an outbreak in a SCC turns into a large outbreak, its final size is independent of its starting point [18, 76]. For a weakly connected network, we can built upon this intuition by defining $f_{i\to j}$ as the fraction of hosts infected in node $j \in \mathcal{S}_l$ during a large outbreak

that also occurred in node $i \in \mathcal{S}_m$. Note that the notation has a different interpretation in this case compared to the previous ones. The outbreak may have started in a different node, $k \in \mathcal{S}_n$. Given that node i experiences a large outbreak, $f_{i\to j}$ is the relative size of the large outbreak in node j. In the EPN framework, it is calculated as

$$f_{i\to j} = 1 - H_{j\leftarrow k}(\mathbf{1}) \quad \forall\, k \in \mathcal{S}_i \tag{4.32}$$

4.5 1D LATTICE WITH DIRECTIONAL COUPLING

We consider a 1D lattice with M nodes and unidirectional coupling (see fig. 4.5). Each node in the lattice has N hosts. Let R_0 be the basic reproduction number associated with each node in the lattice and $\rho_{i,i+1}$ be the directional coupling from node i to $i+1$. Further, we assume that the rate of infectious contact is β in all nodes and that infectious period is identically and exponentially distributed for each node, i.e., $F_2(\tau) \sim \text{Exp}(1/\mathbb{E}[\tau])$. We now define the PGF $G_i^{in}(\boldsymbol{x})$ and $G_i^{out}(\boldsymbol{x})$ for the distribution of in-degree and out-degree in the EPN in the limit $N \to \infty$,

$$\begin{aligned} G_i^{in}(x_i, x_{i-1}) &= \begin{cases} \exp\left[R_0(x_i - 1)\right] & i = 1, \\ \exp\left[R_0\left\{(x_i - 1) + \rho_{i-1,i}(x_{i-1} - 1)\right\}\right] & i \geq 2. \end{cases} \\ G_i^{out}(y_i, y_{i+1}) &= \frac{1}{1 - R_0\left[(y_i - 1) + \rho_{i,i+1}(y_{i+1} - 1)\right]} \end{aligned} \tag{4.33}$$

and their product $G_i(x_i, x_{i-1}, y_i, y_{i+1})$ is

$$G_i(x_i, x_{i+1}, y_i, y_{i+1}) = \begin{cases} \dfrac{\exp\left[R_0(x_i - 1)\right]}{1 - R_0\left[(y_i - 1) + \rho_{i,i+1}(y_{i+1} - 1)\right]} & i = 1, \\ \dfrac{\exp\left[R_0\left\{(x_i - 1) + \rho_{i-1,i}(x_{i-1} - 1)\right\}\right]}{1 - R_0\left[(y_i - 1) + \rho_{i,i+1}(y_{i+1} - 1)\right]} & i \geq 2. \end{cases} \tag{4.34}$$

We now apply the formalism presented in the previous section to calculate the distribution

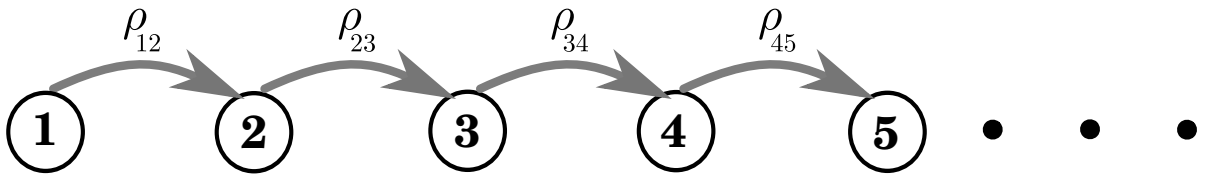


Figure 4.5: A 1D lattice with unidirectional coupling.

of outbreak sizes, the epidemic threshold and the relative size of large outbreak. If infection enters the system at node i, the distribution of outbreak sizes in that node is given by,

$$H_{i\to i}(y) = yG_i^{out}(H_{i\to i}(y), 1) \tag{4.35}$$

whose solution is the PGF that was derived earlier for the linear birth-death process (eq. 1.22):

$$H_{i\to i}(y) = \frac{R_0 + 1 - \sqrt{(R_0+1)^2 - 4R_0 y}}{2R_0} \tag{4.36}$$

The distribution of outbreak sizes in node j when infection starts in node i $(i < j)$ is given by

$$H_{i\to j}(y) = G_i^{out}(H_{i\to j}(y), H_{i+1\to j}(y)) \tag{4.37}$$

Using the PGF from eq. 4.37 and assuming equal coupling ρ between all consecutive nodes in the lattice, we calculate the probability of spillover $P_{1\to n}$, the probability of a large outbreak $Q_{1\to n}$ and the probability of a small outbreak $U_{1\to n}$ in node n given that the outbreak starts in node 1. The analytical results are plotted in figure 4.6 along with results from stochastic simulations. The epidemic threshold for each node in the network is $R_0 = 1$. This is because there are no strongly connected components and with each node having the same value of the basic reproduction number, a large outbreak can only occur if R_0 exceeds 1. The probability of spillover decreases because small outbreaks die out exponentially with node distance. The probability of a large outbreak increases because an outbreak that was small in node n can potentially start a large outbreak in node $n+1$. This probability levels

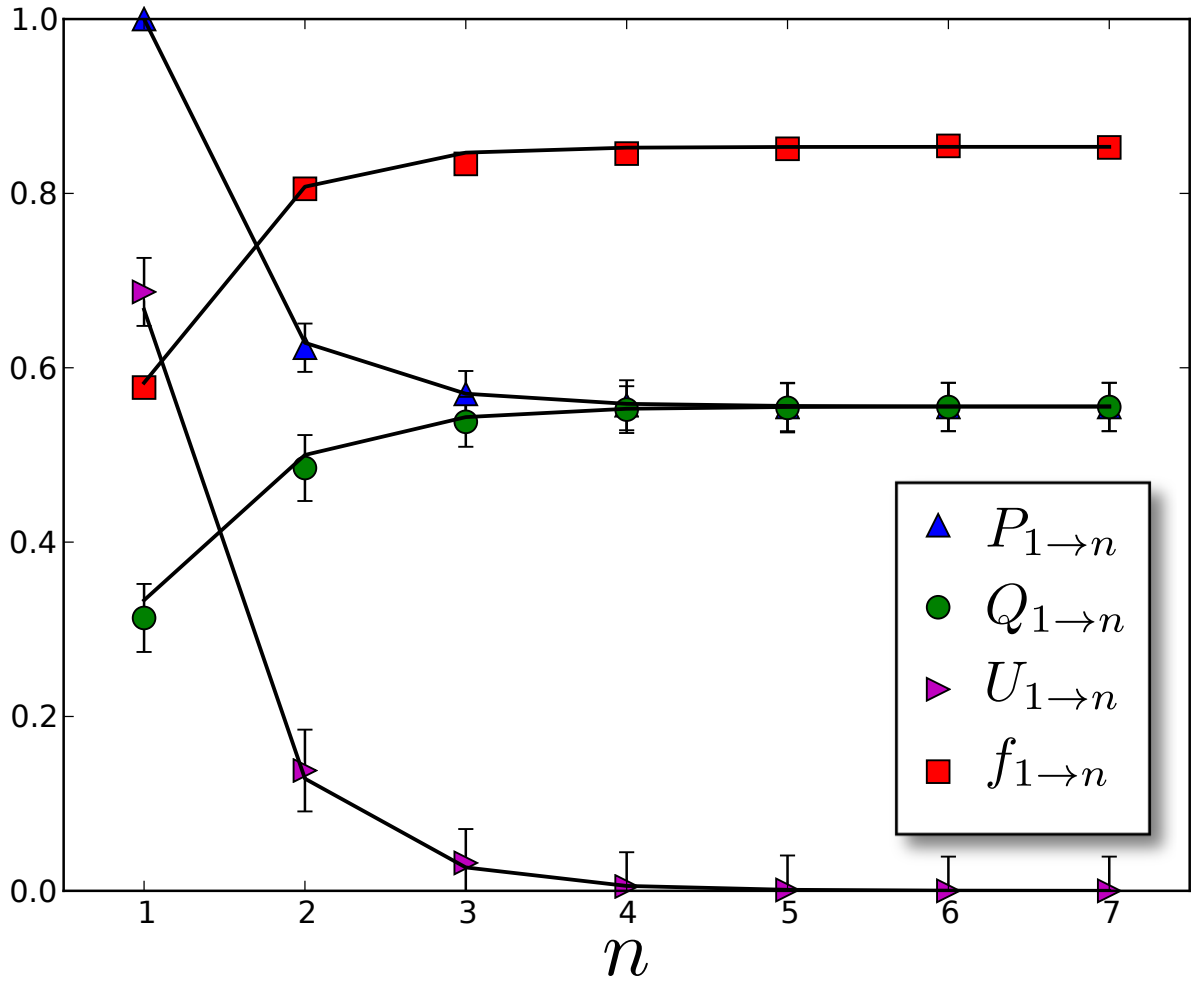


Figure 4.6: The analytical results (solid lines) obtained by applying the formalism to a 1D lattice and comparison with 1000 stochastic simulations (points with errorbars). The parameters in the model are as follows: $N = 1000, R_0 = 1.5, \rho_{i,i+1} = 0.5$. The Outbreak first starts in node 1. The probability of a small outbreak decreases with node distance whereas the probability of a large outbreak increases. The probability of spillover and that of a large outbreak converge for large n which implies that only large outbreaks propagate on the lattice whereas the small ones die out exponentially with the node distance.

off with increasing n because small outbreaks do not survive for longer node distances and only large outbreaks propagate on the lattice [59]. Thus $P_{1\to n}$ and $Q_{1\to n}$ converge to the same constant value and $U_{1\to n}$ converges to 0, all exponentially with n.

The value to which $P_{1\to n}$ and $Q_{1\to n}$ converge can be calculated as follows. From eq. 4.37, the probability of extinction can be calculated as

$$\pi_{1\to n} = G_1^{out}(\pi_{1\to n}, \pi_{2\to n}) \tag{4.38}$$

In the limit of $n \to \infty$, the probability of extinction in node n will be independent of the starting point. Thus, $\pi_{1\to n} = \pi_{2\to n} = \ldots \pi_{j\to n} = \pi_\infty$. The equation now reduces to

$$\pi_\infty = G_1^{out}(\pi_\infty, \pi_\infty) \tag{4.39}$$

whose solution can be obtained analytically as follows

$$\pi_\infty = \min\left(1, \frac{1}{R_0(1+\rho)}\right) \tag{4.40}$$

Thus, an epidemic propagating on an infinite lattice has a lower epidemic threshold, i.e.,

$$R_0(1+\rho) = 1 \tag{4.41}$$

which is distinct from the epidemic threshold if the lattice has a finite number of nodes, i.e., $R_0 = 1$.

Interestingly, if $R_0 < 1$ and $R_0(1+\rho) > 1$, the infection will continue to propagate on an infinite lattice but the size of the outbreaks in individual nodes of the lattice will be small.

The relative final size of a large outbreak can be obtained using eq. 4.31, 4.32 and 4.33. Let $f_{1\to n}$ be the relative size of the large outbreak in node n given that there occurs a large outbreak in node 1. $f_{1\to n}$ can be solved using the following equation

$$1 - f_{1\to n} = \begin{cases} \exp\left[-R_0 f_{1\to n}\right] & n = 1, \\ \exp\left[-R_0\left(f_{1\to n} + \rho f_{1\to n-1}\right)\right] & n \geq 2. \end{cases}$$

Numerical solutions of eq. 4.42 are plotted in figure 4.6 and compared with results of stochastic simulations. The error bars for $f_{1\to n}$ in the plot are smaller than the markers. For $R_0 > 1$, the size of the outbreak increases with n and converges to a constant value in the limit of large n. The limit can be calculated by replacing $f_{1\to n}$ and $f_{1\to n-1}$ in eq. 4.42 by f_∞, which implies that for large n the size of the large outbreak in nodes n and $n-1$ will be equal. f_∞ can be solved as the solution to the following equation

$$1 - f_\infty = \exp\left[-R_0(1+\rho)f_\infty\right] \tag{4.42}$$

This is consistent with the fact that $R_0(1+\rho)$ is the effective basic reproduction number for the infinite lattice.

An interesting ambiguity arises if $R_0 < 1$ and $R_0(1+\rho) > 1$. In this case a large outbreak cannot occur in a lattice with a finite number of nodes. But for an infinite lattice, eq. 4.42 suggests otherwise for large n. The ambiguity is resolved by the fact that the large outbreak only occurs in a node that is infinitely far away from the first node. The presence of two infinities (population size and lattice length) is at the source of this indeterminacy.

4.6 Risk Maps

In this example, we demonstrate the utility of the presented formalism by constructing risk maps for a contrived metapopulation network (figure 4.7). This network can be imagined to be the livestock supply chain with node 5 as the slaughterhouse, nodes 4,7 and 9 being the feedlots and others being farms or cattle markets. The calculations do not assume anything about the physical picture. They only rely on the structure of the network and the model of disease dynamics. In this example we fix the intra-population Basic Reproduction Number

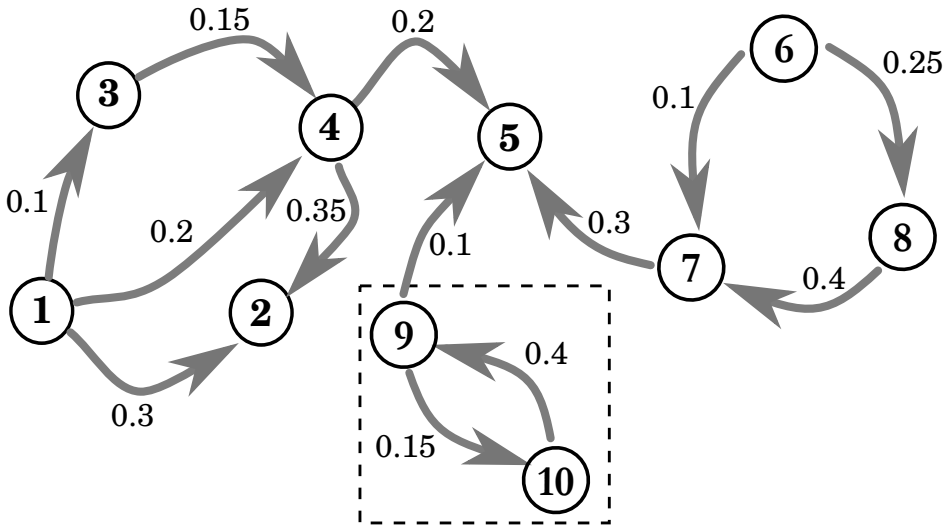


Figure 4.7: An example of a weakly connected network to demonstrate the utility of risk maps. This can be an example of a livestock supply chain with node 5 as the slaughterhouse. Nodes 9 and 10 constitute a strong connected subgraph in the network. The risk map is a tool to evaluate the risk of a large outbreak in different nodes of the network.

to be the same as R_0, the infectious period distribution to be exponential and the coupling matrix specified below

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.3 & 0.1 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.15 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.35 & 0 & 1 & 0.2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.1 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 1 & 0.15 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.4 & 1 \end{bmatrix} \tag{4.43}$$

The $\boldsymbol{\rho}$ matrix can be specified from data if the rate of migration is known [17]. Proceeding in the manner similar to the previous section, we write down the generating functions $G_i(\boldsymbol{x}, \boldsymbol{y})$ and $H_{i\to j}(y)$ for each set of (i, j).

For $R_0 = 1.1$, we obtain the following matrices whose $(i,j)^{th}$ elements represent $P_{i\to j}$ and $Q_{i\to j}$ respectively.

$$[P_{i\to j}] = \begin{bmatrix} 1 & 0.51 & 0.31 & 0.42 & 0.32 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.28 & 1 & 0.36 & 0.26 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.48 & 0 & 1 & 0.40 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.35 & 1 & 0.41 & 0.43 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.45 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.38 & 0 & 0.50 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.36 & 0 & 0 & 0 & 1 & 0.36 \\ 0 & 0 & 0 & 0 & 0.35 & 0 & 0 & 0 & 0.50 & 1 \end{bmatrix} \tag{4.44}$$

$$[Q_{i\to j}] = \begin{bmatrix} 0.09 & 0.30 & 0.14 & 0.22 & 0.25 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.09 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.21 & 0.09 & 0.16 & 0.20 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.21 & 0 & 0.09 & 0.18 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.09 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.30 & 0.09 & 0.27 & 0.19 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.20 & 0 & 0.09 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.29 & 0 & 0.22 & 0.09 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.26 & 0 & 0 & 0 & 0.24 & 0.24 \\ 0 & 0 & 0 & 0 & 0.31 & 0 & 0 & 0 & 0.31 & 0.31 \end{bmatrix}$$

The difference between the two matrices $[U_{i\to j}]$ gives the probability of a small outbreak in

node j given that infection starts in node i.

$$[U_{i\to j}] = \begin{bmatrix} 0.91 & 0.20 & 0.17 & 0.20 & 0.06 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.91 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.07 & 0.91 & 0.20 & 0.06 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.27 & 0 & 0.91 & 0.22 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.91 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.05 & 0.91 & 0.14 & 0.24 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.25 & 0 & 0.91 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.09 & 0 & 0.28 & 0.91 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.11 & 0 & 0 & 0 & 0.76 & 0.12 \\ 0 & 0 & 0 & 0 & 0.04 & 0 & 0 & 0 & 0.19 & 0.69 \end{bmatrix} \tag{4.45}$$

In this example, we consider the process of assigning the probability of infection origin to each node in the network given that disease is detected in a particular node, such as 5. We can calculate the probability using Bayes' theorem assuming a uniform prior

$$P_{k\,|\,i} = \frac{P_{i\to k}}{\sum_{j=1}^{M} P_{j\to k}} \tag{4.46}$$

Similarly, we can define $Q_{k\,|\,i}$ and $U_{k\,|\,i}$ conditioning on the observed outbreak being large or small. For the presented example, the vector $P_{k\,|\,5}$ is calculated as

$$P_{k\,|\,5} \approx [0.08, 0, 0.07, 0.10, 0.26, 0.09, 0.12, 0.10, 0.09, 0.09] \tag{4.47}$$

The risk map for $P_{k\,|\,5}$ is shown in figure 4.8 with a gray coloring scheme.

Similar calculation for node 2

$$P_{k\,|\,2} \approx [0.22, 0.44, 0.12, 0.21, 0, 0, 0, 0, 0] \tag{4.48}$$

The most likely point of origin is the node where the disease is first observed. But the calculation also assign meaningful probabilities to other nodes in the network. This information can be used to design surveillance strategies for minimizing the risk of global spread as well as to plan control strategies in the event of an outbreak.

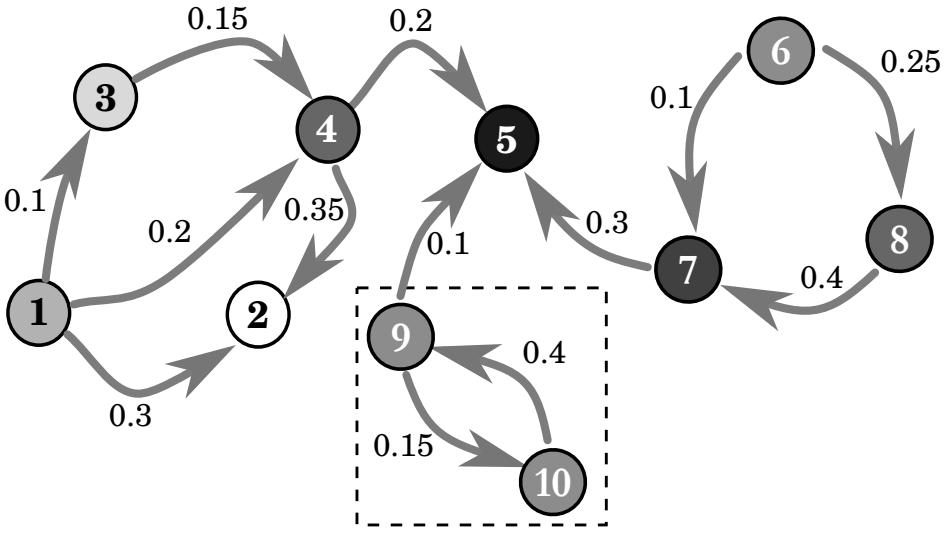


Figure 4.8: The metapopulation network of figure 4.7 colored according to the probability $P_{k\,|\,5}$ (eq. 4.47), i.e., of a node being the starting point of an outbreak that is detected in node 5. The coloring scheme does not represent the probabilities but rather the ordering of the probabilities. Darker shade signifies a higher probability.

4.7 Discussion

In this work we have presented a methodology for analytically describing disease spread in metapopulation networks where host population has two levels of structure. Such networks naturally arise in many contexts, and could be constructed from migration patterns of hosts among subpopulations. In the limit of large population, the formalism provides analytical solutions that agree with stochastic simulations. Although the results are presented for a fully mixed model, the formalism is general and applicable to metapopulation networks with heterogeneous contact structure among hosts [59, 75]. For a large class of such models, expensive simulations can be avoided in favor of analytical results.

One of the most important pieces of information that can be inferred from our analysis is the probability of a node in the metapopulation network becoming infected at some time during an outbreak. From the perspective of public health policy, these results can be

applied to assess the vulnerability of certain network structures towards disease outbreaks such as in the case of farm networks or livestock supply chains.

APPENDIX A

Probability generating functions

A.1 Generating functions

A generating function is a power series representation for a sequence that converges in a given domain D.

$$\mathcal{A}(x) = \sum_{n=0}^{\infty} a_n x^n, \quad |x| \in D$$

$$a_n = \frac{1}{n!}\mathcal{A}^{(n)}(0)$$

The function $\mathcal{A}(x)$ serves as a 'generator' for the sequence $\{a_n\}$. For example,

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \quad |x| < 1$$

A.2 Probability generating functions (PGF)

For a discrete random variable $X > 0$, the probability generating function (PGF) is a power series representation of its probability mass function, i.e.,

$$G(s) = \mathbb{E}[s^X] = \sum_{n=0}^{\infty} p_n s^n \tag{A.1}$$

from which p_n can be sampled in the following manner:

$$p_n = \frac{1}{n!}G^{(n)}(0) \tag{A.2}$$

By virtue of being a PGF, $G(s)$ is endowed with the following properties:

1. $G(1) = 1$
2. $G^{(n)}(0) > 0$ for all n.
3. $G(s)$ is a monotonically increasing and bounded in $[0, 1]$.

Moments of the distribution can be obtained by taking derivatives of the PGF, i.e.,

$$\mathbb{E}[X] = \sum_{n=0}^{\infty} np_n = G'(1)$$

A.3 PGF FOR JOINT DISTRIBUTION

For two random variables X and Y, the joint PGF is given by

$$G_{X,Y}(u, v) = \mathbb{E}[u^X v^Y] = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} p_{m,n} u^m v^n$$

$$p_{m,n} = \frac{G_{X,Y}^{(m,n)}(0, 0)}{m!n!}$$

If X and Y are independent, the PGF for the joint distribution is the product of the individual PGFs, i.e.,

$$G_{X,Y}(u, v) = G_X(u) G_Y(v)$$

And the same holds true for the sum of independent random variables, i.e.,

$$G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X] \cdot \mathbb{E}[s^Y] = G_X(s) G_Y(s)$$

For N i.i.d random variables $X_i, 1 \leq i \leq N$, let $S_N = \sum_{i=0}^{N} X_i$

$$G_{S_N}(s) = [G_{X_1}(s)]^N$$

If N is also a random variable,

$$\begin{aligned} G_{S_N}(s) &= \mathbb{E}[s^{S_N}] = \mathbb{E}[s^{\sum_{i=1}^{N} X_i}] \\ &= \mathbb{E}\left[\mathbb{E}[s^{\sum_{i=1}^{N} X_i}|N]\right] = \mathbb{E}[G_{X_1}(s)^N] \\ &= G_N(G_{X_1}(s)) \end{aligned} \tag{A.3}$$

For more details on the use and properties of generating functions, the reader is encouraged to explore references [12, 84, 85].

A.4 Sampling from a PGF

As described in section A.1, the probability p_n can be obtained by taking derivatives of the PGF, i.e.,

$$p_n = \frac{1}{n!} G^{(n)}(0) \tag{A.4}$$

For many generating functions that are encountered in applied problems it is not feasible to evaluate higher order derivatives analytically or have a power series representation. In such cases, we resort to numerical methods. Since, numerical derivates (e.g., based on finite differences) can be inaccurate, a convenient alternative is to use the Cauchy integral formula,

$$p_n = \frac{1}{2\pi i} \oint \frac{G(x)}{x^{n+1}}\, dx \tag{A.5}$$

which can be evaluated with a high degree of precision using numerical methods. The integral in the Cauchy formula is over a unit circle in the complex plane.

A.5 PGF of a homogeneous Poisson process

A homogeneous Poisson process is defined by the following rate equation

$$X \xrightarrow{\lambda} X + 1 \tag{A.6}$$

The random variable $X(t)$ is a counter for the number of events that have occurred by time t. Events occur with a constant probabilistic rate λ, i.e.,

$$\mathbb{P}[X(t + dt) = n + 1 \mid X(t) = n] = \lambda dt$$

Increments in the process are stationary and independent, i.e., for $s \leq t$ and $i \leq j$,

$$\mathbb{P}[X(t) = j | X(s) = i] \;=\; \mathbb{P}[X(t - s) = j | X(0) = i] \;=\; \mathbb{P}[X(t - s) = j - i]$$

The master equation for $\mathbb{P}[X(t) = n] \equiv p_n$ can be written as:

$$\frac{dp_n}{dt} = \lambda(p_{n-1} - p_n) \tag{A.7}$$

Let $G(x, t) = \sum_{n=0}^{\infty} p_n(t)x^n$ be the PGF of $X(t)$. We multiply the master equation by x^n and sum over n to obtain the following PDE

$$\frac{\partial G}{\partial t} = \lambda(x - 1)G \tag{A.8}$$

The solution to the PDE after using the boundary condition $G(1, t) = 1$ is

$$G(x, t) = e^{\lambda t(x-1)} = \sum_{k=0}^{\infty} \frac{e^{-\lambda t}(\lambda t)^k}{k!} x^k \tag{A.9}$$

APPENDIX B

Galton-Watson Branching Process

The Galton-Watson branching process is a discrete time Markov process which models the growth of a population where individuals produce offspring independent of each other and with a given distribution [11].

B.1 Distribution of individuals in n^{th} generation

Let Y_n be the total population in the n^{th} generation and $X_{n,i}$ $(1 \leq i \leq Y_{n-1})$ be the number of offspring of each individual of the $(n-1)^{th}$ generation. Y_n can be expressed in

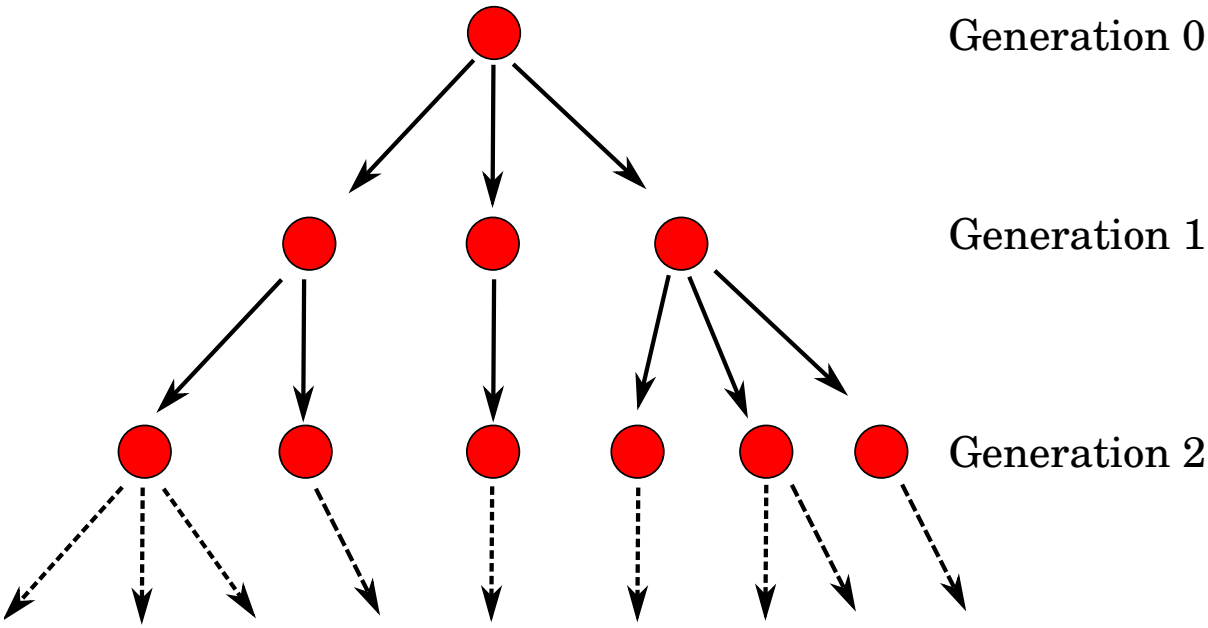


Figure B.1: A schematic of the Galton-Watson branching process.

terms on Y_{n-1} and $X_{n,i}$, i.e.,

$$Y_n = \sum_{i=0}^{Y_{n-1}} X_{n,i}$$

Let $G_X(s)$ be the PGF for the offspring distribution and $G_{Y_n}(s)$ be the PGF for the population in n^{th} generation. From eq. A.3, we obtain

$$G_{Y_n}(s) = G_{Y_{n-1}}(G_X(s)) \tag{B.1}$$

Since the process starts with 1 individual in the 0^{th} generation,

$$G_{Y_0}(s) = s \tag{B.2}$$

which leads to

$$G_{Y_1}(s) = G_X(s) \tag{B.3}$$

and

$$G_{Y_n}(s) = \underbrace{G_X(G_X(\ldots))}_{n \text{ compositions}} \tag{B.4}$$

B.2 Distribution of total individuals ever born

Let $Z = \sum_{i=0}^{\infty} Y_i$ be the total individuals ever born in the Galton-Watson branching process. Z can be written as

$$Z = 1 + \sum_{i=0}^{Y_1} Z_i \tag{B.5}$$

where Z_i is the sum of all individuals that originate from i^{th} individual in the 1^{st} generation. Z and $Z_i, 1 \leq i \leq Y_1$ are i.i.d. and thus, will have the same PGF. From eq. B.5, we can write down the following PGF equation:

$$\begin{aligned} G_Z(s) &= sG_{Y_1}(G_Z(s)) \\ &= sG_X(G_Z(s)) \end{aligned} \tag{B.6}$$

The average size of the population is given by $G'_Z(1)$:

$$G'_Z(1) = \frac{1}{1 - G'_X(1)} \tag{B.7}$$

$G'_X(1)$ is the Expected value of the offspring distribution. From eq. B.7, $G'_X(1) = 1$ is a critical point for the branching process. If $G'_X(1) < 1$, the expected value of Z is finite, i.e., the process goes extinct in finite generations. This means that if each individual produces less than 1 offspring on average, the process will always go towards extinction. This is the definition of a *subcritical* process. On the other hand, if each individual produces more than 1 offspring on average, then the process will diverge. Interestingly, a supercritical process ($G'_X(1) > 1$) can also go extinct in finite generations due to randomness in the offspring distribution. This is discussed in the next section.

B.3 Probability of extinction

A keys property of a PGF is that it always represents the probability mass associated with the finite state space. For random variables that are restricted to a finite state space, this does not affect the properties of their generating functions, such as the fact that $G(1)$ is always equal to 1. But for random variables defined on an infinite state space such as the Galton-Watson branching process, their generating functions become defective, i.e, $G_Z(1) \leq 1$. If $G_Z(1) = 1$, it implies that the entire distribution of the process is contained in a finite state space whereas if $G_Z(1) < 1$ then it implies that the defective probability mass $1 - G_Z(1)$ is concentrated at ∞. By these arguments, $G_Z(1)$ is the probability of extinction of a branching process. Let this probability be represented by π. From eq. B.6, π is the solution of the following self-consistent equation:

$$\pi = G_X(\pi) \tag{B.8}$$

If $G'_X(1) < 1$, then $\pi = 1$ is the only solution to eq. B.8. For $G'_X(1) > 1$, there will be at least two positive solutions to the equation B.8, one of which will be $\pi = 1$. The probability of extinction in that case is the smallest positive solution of the equation. If $G'_X(1) = 1$, probability of extinction is 1 if the variance of the offspring distribution is positive and 0 otherwise. For technical details surrounding these results, please consult [85].

APPENDIX C

Appendix to chapter 2

C.1 Derivation of generating function for the BDI process

As described in the main text, the statistics of the BDI process are mathematically equivalent to the statistics of the $M/G/\infty$ queue. From [33], the number of customers served in the busy period of an $M/G/\infty$ queue with arrival rate λ and service time distribution $U(s)$ is generated by the following PGF

$$\begin{aligned} G(x) &= 1 - \frac{1}{\lambda Q(x)} \\ Q(x) &= \int_0^\infty \exp\left[-\lambda t + \lambda x \int_0^t U(s)ds\right] dt \end{aligned} \tag{C.1}$$

The number of customers served corresponds to the number of micro-outbreaks in the BDI process that occur on overlapping time-scales. The distribution $U(s)$ corresponds to the duration of a micro-outbreak, i.e., the BD process. The joint distribution of duration T and size $R(T)$ of an outbreak in the BD process is generated by

$$\begin{aligned} F(0, y; s) &= \sum_{n\geq 1} \mathbb{P}\big[T \leq s, R(T) = n\big]\, y^n \\ &= \frac{\Lambda_0\Lambda_1\left(1 - e^{-\alpha(\Lambda_1 - \Lambda_0)s}\right)}{\Lambda_1 - \Lambda_0 e^{-\alpha(\Lambda_1 - \Lambda_0)s}} \end{aligned} \tag{C.2}$$

where $\Lambda_0(y)$ and $\Lambda_1(y)$ are roots of the following quadratic equation such that $0 < \Lambda_0 < 1 < \Lambda_1$.

$$\alpha w^2 - (\alpha + 1)\, w + y = 0 \tag{C.3}$$

Substituting $\nu\alpha$ for λ and $F(0, y; s)$ for $U(s)$ in eq. C.1 and simplifying the integral, we obtain the PGF for the joint distribution of number of micro-outbreaks and outbreak size in the BDI process.

$$G(x, y) = 1 - \frac{1}{\nu} \frac{\Lambda_1\, z^a\, (1 - z)^b}{\int_z^1 r^{a-1}(1 - r)^{b-1} dr} \tag{C.4}$$

where

$$z = 1 - \frac{\Lambda_0}{\Lambda_1}, \quad a = 1 - \nu x, \quad b = \nu \left(\frac{1 - \Lambda_0 x}{\Lambda_1 - \Lambda_0} \right)$$
$$\Lambda_0, \Lambda_1 = \frac{(\alpha + 1) \mp \sqrt{(\alpha + 1)^2 - 4\alpha y}}{2\alpha}$$

C.2 Survival function for $\nu > 1$

To calculate the survival function for $\alpha = 1, \nu > 1$, we repeat the calculation of the previous section but limit the integration in eq. C.1 to a finite (but large) t rather than ∞. In doing so, the PGF $G(x, y; t)$ reflects the distribution for those outbreaks that end before time t.

$$G(x, y; t) = 1 - \frac{1}{\nu} \frac{\Lambda_1\, z_0^a\, (1 - z_0)^b}{\int_{z_0}^{z_t} r^{a-1}(1 - r)^{b-1} dr} \tag{C.6}$$

where

$$z_t = 1 - \frac{\Lambda_0}{\Lambda_1} e^{-(\Lambda_1 - \Lambda_0)t} \tag{C.7}$$

and α is set to 1 for Λ_0 and Λ_1. The distribution function for the duration T is the total probability contained in the PGF, i.e.,

$$\mathbb{P}\big[T < t\big] = \lim_{(x,y)\to(1,1)} G(x, y; t) \tag{C.8}$$

In the limit $(x, y) \to (1, 1)$,

$$
\begin{aligned}
&a \to 1-\nu, \quad b \to \frac{\nu}{2}, \quad \Lambda_0, \Lambda_1 \to 1, \\
&z_t \to 2\,(1+t)\sqrt{1-y}, \\
&\int_{z_0}^{z_t} r^{a-1}(1-r)^{b-1} dr \to \frac{z_t^a - z_0^a}{a}
\end{aligned}
$$

Taking the limit and simplifying the expression, we obtain

$$
\mathbb{P}\big[T < t\big] \sim \frac{\nu^{-1} - (1+t)^{1-\nu}}{1-(1+t)^{1-\nu}} \tag{C.9}
$$

As $t \to \infty$, the probability converges to ν^{-1} which is the probability that the outbreak has a finite size. With probability $1-\nu^{-1}$ the outbreak persists indefinitely. The survival function $P(t)$ is thus defined for finite size outbreaks,

$$
\begin{aligned}
P(t) &= \mathbb{P}\big[\, t < T < \infty\big] \\
&\sim \frac{1-\nu^{-1}}{(1+t)^{\nu-1}-1} \\
&\sim \frac{1}{t^{\nu-1}} \quad \text{for large } t
\end{aligned} \tag{C.10}
$$

The average duration can be calculated as

$$
\langle t \rangle_{pl} \;\sim\; -\int_0^{T_c} t \frac{dP}{dt} dt \;\sim\; \begin{cases} T_c^{2-\nu} & \nu \in (1,\infty)\backslash\{2\}, \\ \log T_c & \nu = 2. \end{cases} \tag{C.11}
$$

where T_c is a cutoff timescale in the BDI process with finite system size. We would like to note that the technique of using the survival function is also applicable when $\nu < 1$ and it yields the desired scaling laws (eq. 2.54) when applied.

APPENDIX D

Appendix to chapter 3

D.1 Moments of first passage time

The distribution of first passage time reported in the main text:

$$\begin{aligned}\mathbb{P}[T \leq t] &= \mathbb{P}[Z_{h,p}(t) > 0] \\ &= 1 - \mathcal{G}_a(1,1,0;t) \\ &= 1 - \frac{v_0(v_1-1)+v_1(1-v_0)e^{-\beta_{aa}(v_1-v_0)t}}{(v_1-1)+(1-v_0)e^{-\beta_{aa}(v_1-v_0)t}}\end{aligned} \tag{D.1}$$

Here we report the calculation of moments for the first passage time distribution:

$$\begin{aligned}\mathbb{E}[T^n \,|\, T<\infty] &= \frac{\mathbb{E}[T^n \mathbf{1}_{\{T<\infty\}}]}{\mathrm{P}[T<\infty]} \\ &= \frac{n\int_0^\infty t^{n-1}\,\mathrm{P}[\,t<T<\infty]\,dt}{\mathrm{P}[T<\infty]} \\ &= n(v_1-v_0)\int_0^\infty \frac{t^{n-1}e^{-\beta_{aa}(v_1-v_0)t}}{(v_1-1)+(1-v_0)e^{-\beta_{aa}(v_1-v_0)t}}dt\end{aligned}$$

where $\mathbf{1}_{\{T<\infty\}}$ is the indicator function that takes the value 1 for $T < \infty$. Letting $c = v_1 - 1, d = 1 - v_0$ and $k = \beta_{aa}(v_1 - v_0)$, we can rewrite this as follows:

$$\begin{aligned}\mathbb{E}[T^n \,|\, T<\infty] &= \frac{nk}{\beta_{aa}}\int_0^\infty \frac{t^{n-1}e^{-kt}}{c+de^{-kt}}\,dt \\ &= \frac{-n!}{\beta_{aa}k^{n-1}d}\,\mathrm{Li}_n\left(\frac{-d}{c}\right) \\ &= \frac{n!}{(\beta_{aa})^n\,(v_0-1)\,(v_1-v_0)^{n-1}}\mathrm{Li}_n\left(\frac{v_0-1}{v_1-1}\right)\end{aligned} \tag{D.2}$$

where $\mathrm{Li}_n(z)$ is the polylogarithm function of order n. The conditional expected value of the first passage time is obtained by setting $n = 1$ in this expression:

$$\mathbb{E}[T \,|\, T < \infty] = \frac{1}{\beta_{aa}\,(1 - v_0)} \log\left(\frac{v_1 - v_0}{v_1 - 1}\right) \tag{D.3}$$

Our result is for all moments of the distribution; to our knowledge, only the first moment has been reported earlier in [52] (in the context of a problem in population genetics).

D.2 Finite-size corrections to probability of spillover

The probability of spillover, as calculated in eq. (3.15), is valid only in the limit of $N_a, N_h \to \infty$. Deviations from this result are expected for finite system sizes, which we report here. We make the assumption that the linear birth-death process provides accurate statistics of small outbreaks in finite size systems. This is a valid assumption since small outbreaks are $o(N)$ and thus, their distribution is independent of the total system size provided $N \gg 1$. Using, eq. (3.17), the assumption is stated mathematically as follows

$$\mathrm{P}^N[\text{spill} \,|\, \text{small outbreak}] = \begin{cases} 1 - \mathcal{R}_0^{aa} v_0 & \text{if } \mathcal{R}_0^{aa} \leq 1, \\ 1 - v_0 & \text{if } \mathcal{R}_0^{aa} > 1. \end{cases} \tag{D.4a}$$

$$\mathrm{P}^N[\text{small outbreak}] = \frac{1}{\mathcal{R}_0^{aa}} \tag{D.4b}$$

Now we calculate the finite size equivalent of $\mathrm{P}[\text{spill} \,|\, \text{large outbreak}]$ using the hazard function. For this calculation, we ignore the fluctuations around the mean and assume that the animal epidemic obeys the deterministic SIR. Before the first primary infection, the entire human population is susceptible and thus $S_{h,1}(t) = \nu N_h$.

$$\begin{aligned} \mathrm{P}^N[\text{spill} \,|\, \text{large outbreak}] &= 1 - \exp\left\{-\int_0^\infty \frac{\beta_{ah} S_{h,1} I_a}{N_a} dt\right\} \\ &= 1 - \exp\left\{-N_a \mathcal{R}_0^{ah} f_a\right\} \end{aligned} \tag{D.5a}$$

where

$$f_a = \lim_{N_a \to \infty} \frac{\mathbb{E}[R_a(\infty)]}{N_a} \tag{D.5b}$$

is obtained by solving the final size equation for a simple SIR

$$1 - f_a = e^{-\mathcal{R}_0^{aa} f_a} \tag{D.5c}$$

From eq. (D.5a), $\mathrm{P}^N[\text{spill} \,|\, \text{large outbreak}] \to 1$ as $N_a \to \infty$ and this agrees with the large system size limit. Using the law of total probability, we now arrive at the probability of spillover with finite size corrections.

$$\begin{aligned} \mathrm{P}^N[\text{spill}; \mathcal{R}_0^{aa} \leq 1] &= 1 - v_0 \\ \mathrm{P}^N[\text{spill}; \mathcal{R}_0^{aa} > 1] &= 1 - v_0 - \left(1 - \frac{1}{\mathcal{R}_0^{aa}}\right) \exp\left\{-N_a \mathcal{R}_0^{ah} f_a\right\} \end{aligned} \tag{D.6}$$

Figure 3.5 shows the comparison of finite size corrections as calculated using eq. (D.6) with stochastic simulations.

In the limit of vanishingly small $\mathcal{R}_0^{ah}$, it is important to consider the limit of $\mathcal{R}_0^{ah} N_a$ as $N_a \to \infty$. Let $\xi = \mathcal{R}_0^{ah} N_a$. The probability of spillover presented in Figure 3.4 assumes the limit of $\xi \to \infty$. For finite sizes, the probability of spillover simplifies to

$$\begin{aligned} \lim_{\substack{\mathcal{R}_0^{ah} \to 0 \\ N_a \to \infty}} \mathrm{P}^N[\text{spill}; \mathcal{R}_0^{aa} \leq 1] &= 0 \\ \lim_{\substack{\mathcal{R}_0^{ah} \to 0 \\ N_a \to \infty}} \mathrm{P}^N[\text{spill}; \mathcal{R}_0^{aa} > 1] &= \left(1 - \frac{1}{\mathcal{R}_0^{aa}}\right) \cdot [1 - \exp\{-\xi f_a\}] \end{aligned} \tag{D.7}$$

Thus, depending on the value of ξ, the limiting value for the probability of spillover when $\mathcal{R}_0^{aa} > 1$ can assume any value in the range $[0, 1 - 1/\mathcal{R}_0^{aa}]$. Thus, if $\mathcal{R}_0^{aa} \gg 1$ and $\mathcal{R}_0^{ah} \ll 1$, then the probability of spillover is indeterminate if there is no information about the scale of $\mathcal{R}_0^{ah} N_a$.

D.3 Asymptotic scaling near the critical threshold

The scaling of the outbreak sizes near the critical threshold can be investigated through the singularity analysis of the associated generated function $H(z)$ [36]. The dominant singularity ζ of the PGF determines the asymptotic form for $\mathrm{P}(n)$ which is the probability of having an outbreak of size n. If a given PGF can be expanded around the singularity such that

$$H(z) \sim \left(1 - \frac{z}{\zeta}\right)^{\alpha} \tag{D.8}$$

then

$$\mathrm{P}(n) \sim \frac{\zeta^{-n} n^{-\alpha-1}}{\Gamma(-\alpha)}\,, n \to \infty \tag{D.9}$$

where $\alpha \notin \mathbb{Z}_{>0}$. The asymptotic form for $\mathrm{P}(n)$ can be derived by substituting eq. (D.8) in the Cauchy integral formula (eq. (3.28)) and making the following substitution

$$z \mapsto \zeta\left(1 + \frac{t}{n}\right) \tag{D.10}$$

Thus, the singularity determines the exponential factor and the asymptotic form of the generating function determines the power-law exponent. By rescaling the function $H(z) \to H(z\zeta)$, the calculation of the power-law exponent is simplified since the singularity is now located at $z = 1$. We now apply this analysis to the generating function $H_h(z)$.

In the two species metapopulation model of chapter 3, let $\Delta_a = 1 - \mathcal{R}_0^{aa}$ and $\Delta_h = 1 - \mathcal{R}_0^{hh}$ be the distances from the critical thresholds. We first calculate the scaling near the threshold $\mathcal{R}_0^{aa} = 1$, i.e., $|\Delta_a| < |\Delta_h|$ and $|\Delta_a| \ll 1$. We assume that the parameters are such that the singularities of the generating function $H_h(z)$ are far apart. The dominant

singularity near the chosen threshold is given by

$$\begin{aligned}\zeta_1 &= \left(1+\frac{\left(\sqrt{\mathcal{R}_0^{aa}}-1\right)^2}{\mathcal{R}_0^{ah}}\right)\left(1-\mathcal{R}_0^{hh}\frac{\left(\sqrt{\mathcal{R}_0^{aa}}-1\right)^2}{\mathcal{R}_0^{ah}}\right)\\ &= 1+\Delta_h\left(\frac{\Delta_a^2}{4\mathcal{R}_0^{ah}}+\mathcal{O}(\Delta_a^3)\right)\end{aligned} \tag{D.11}$$

The singularity ζ_1 determines the exponential prefactor. To obtain the power-law scaling, the generating function can be analyzed at the critical point ($\mathcal{R}_0^{aa}=1$ in this case) without loss of generality. At the critical point $\zeta_1=1$ and the PGF $H_{h,p}(x)$ simplifies as follows

$$H_{h,p}(z)=\frac{2+\mathcal{R}_0^{ah}(1-z)-\sqrt{\mathcal{R}_0^{ah}(1-z)(4+\mathcal{R}_0^{ah}(1-z))}}{2} \tag{D.12}$$

For further simplification, let $z\hat{H}_{h,s}(z)$ be denoted by $\tilde{H}_{h,s}(z)$. Making the substitution (D.10) and performing a series expansion in fractional powers of $(-t/n)$ gives

$$\tilde{H}_{h,s}(1+t/n)\sim 1+\frac{t}{\Delta_h n} \tag{D.13}$$

Using (3.26), we obtain

$$\begin{aligned}H_h(1+t/n)\sim 1+\frac{\mathcal{R}_0^{ah}}{2\Delta_h}\left(\frac{-t}{n}\right)-\sqrt{\frac{\mathcal{R}_0^{ah}}{\Delta_h}}\left(\frac{-t}{n}\right)^{1/2}\\ -\frac{1}{8}\left(\frac{\mathcal{R}_0^{ah}}{\Delta_h}\right)^{3/2}\left(\frac{-t}{n}\right)^{3/2}\end{aligned} \tag{D.14}$$

By using the Cauchy integral formula on the asymptotic expansion of $H_h(z)$, we obtain the asymptotic probability $P_1(n)$

$$P_1(n)\sim n^{-3/2},\quad \zeta_1=1 \tag{D.15a}$$

at the threshold boundary $\mathcal{R}_0^{aa}=1,\mathcal{R}_0^{hh}\neq 1$. Using the exponential prefactor obtained in eq. (D.11) we arrive at the asymptotic scaling for large n near $\mathcal{R}_0^{aa}=1$:

$$P_1(n)\sim \zeta_1^{-n}n^{-3/2} \tag{D.15b}$$

This scaling of outbreak sizes near $\mathcal{R}_0^{aa} = 1$ is shown in Figure 3.8B. Note that the scaling can be guessed by looking at the leading term in the expansion, which in eq. (D.14) is $(-t/n)^{1/2}$. Similarly, performing the same steps of analysis near the critical point of $\mathcal{R}_0^{hh} = 1$, we obtain

$$P_2(n) \sim \zeta_2^{-n} n^{-3/2} \tag{D.16}$$

where

$$\zeta_2 = 1 + \frac{\Delta_h^2}{4} \tag{D.17}$$

This scaling of outbreak sizes near $\mathcal{R}_0^{hh} = 1$ is shown in Figure 3.8D. Near the multicritical point $\mathcal{R}_0^{aa} = \mathcal{R}_0^{hh} = 1$, the function has a unique singularity if the value of the function $\tilde{H}_{h,s}(z)$ at its singularity ζ_2 coincides with the singularity of the function $H_{h,p}(z)$, i.e.,

$$1 + \frac{\left(\sqrt{\mathcal{R}_0^{aa}} - 1\right)^2}{\mathcal{R}_0^{ah}} = \frac{\mathcal{R}_0^{hh} + 1}{2\mathcal{R}_0^{hh}} \tag{D.18}$$

which simplifies to

$$\Delta_h = \frac{\Delta_a^2}{2\mathcal{R}_0^{ah}} + \mathcal{O}(\Delta_a^3) \tag{D.19}$$

for $\Delta_a, \Delta_h \ll 1$. The unique singularity is given by ζ_2. Thus, the correction to the pure power-law would be ζ_2^{-n}, but only on the curve given by eq. (D.19). Next, we extract the power-law scaling at the threshold. For $\mathcal{R}_0^{aa} = \mathcal{R}_0^{hh} = 1$,

$$\begin{aligned} \tilde{H}_{h,s}(z) &= 1 - \sqrt{1-z} \\ H_{h,p}(z) &= \frac{2 + \mathcal{R}_0^{ah}(1-z) - \sqrt{\mathcal{R}_0^{ah}(1-z)(4 + \mathcal{R}_0^{ah}(1-z))}}{2} \end{aligned} \tag{D.20}$$

whose functional composition yields

$$\begin{aligned} H_h(z) &= H_{h,p}(\tilde{H}_{h,s}(z)) \\ H_h(z) &= \frac{2 + \mathcal{R}_0^{ah}\sqrt{1-z} - \sqrt{\mathcal{R}_0^{ah}\sqrt{1-z}(4 + \mathcal{R}_0^{ah}\sqrt{1-z})}}{2} \end{aligned}$$

Substituting (D.10) and performing a series expansion in fractional powers $(-t/n)$, we obtain the $(-t/n)^{1/4}$ as the leading term. Using the Cauchy integral formula, the asymptotic scaling is given by

$$P_3(n) \sim n^{-5/4}, \quad \zeta_3 = 1 \tag{D.21a}$$

Away from the multicritical threshold but staying on the curve (D.19), the asymptotic form is

$$P_3(n) \sim \zeta_3^{-n} n^{-5/4} \tag{D.21b}$$

where $\zeta_3 = \zeta_2$ as defined in eq. (D.17). This scaling of outbreak sizes at the multicritical point $\mathcal{R}_0^{aa} = 1, \mathcal{R}_0^{hh} = 1$ is shown in Figure 3.8C.

APPENDIX E

Appendix to chapter 4

E.1 Epidemic Percolation Network

The derivation presented here follows [40, 55]. There are M nodes in the metapopulation network. Consider two nodes in this network, i and j, where i and j can be the same node. To obtain one realization of the epidemic percolation network, we shall calculate the probability of transmission in either direction between two randomly chosen host nodes p and q in $\mathcal{P}_i$ and $\mathcal{P}_j$ respectively. Let τ_i and τ_j be the recovery periods of the hosts where $\tau_i \sim F(\tau_i)$ and $\tau_j \sim F(\tau_j)$. Also, we define β_i and β_j as infection contact rate coefficients. For every host edge (p, q), convert the host edge to a directed host edge from p to q with probability $(1 - e^{-\rho_{ij}\beta_i\tau_i/N_j})e^{-\rho_{ji}\beta_j\tau_j/N_i}$, to a directed host edge from j to i with probability $e^{-\rho_{ij}\beta_i\tau_i/N_j}(1 - e^{-\rho_{ji}\beta_j\tau_j/N_i})$, and erase the edge with probability $e^{-\rho_{ij}\beta_i\tau_i/N_j}e^{-\rho_{ji}\beta_j\tau_j/N_i}$. The host edge remains undirected with probability $(1 - e^{-\rho_{ij}\beta_i\tau_i/N_j})(1 - e^{-\rho_{ji}\beta_j\tau_j/N_i})$.

Let $\mathcal{G}_i(\boldsymbol{z})$ be the PGF for the degree distribution in the contact network. Since we are working with a fully mixed system where every host is connected to every other host.

$$\mathcal{G}_i(\boldsymbol{z}) = \prod_{j=1}^{M} (z_j)^{N_j - \delta_{ij}} \tag{E.1}$$

where δ_{ij} is the *kronecker* delta.

Let $g_{ij}(x_j, y_j, u_j \,|\, \beta_i, \beta_j, \tau_i, \tau_j)$ be the conditional PGF for the number of incoming,

outgoing, and undirected host edges incident to p between the host nodes p and q in the *percolation network* where $p \in \mathcal{P}_i$ and $q \in \mathcal{P}_j$.

$$\begin{aligned} g_{ij}(x_j, y_j, u_j \,|\, \beta_i, \beta_j, \tau_i, \tau_j) &= e^{-\rho_{ij}\beta_i\tau_i/N_j - \rho_{ji}\beta_j\tau_j/N_i} \\ &+ e^{-\rho_{ij}\beta_i\tau_i/N_j}\left(1 - e^{-\rho_{ji}\beta_j\tau_j/N_i}\right) x_j \\ &+ \left(1 - e^{-\rho_{ij}\beta_i\tau_i/N_j}\right) e^{-\rho_{ji}\beta_j\tau_j/N_i} y_j \qquad \text{(E.2)} \\ &+ \left(1 - e^{-\rho_{ij}\beta_i\tau_i/N_j}\right)\left(1 - e^{-\rho_{ji}\beta_j\tau_j/N_i}\right) u_j \end{aligned}$$

(E.3)

Before taking the limit on N, it is convenient to expand the terms to 1^{st} order.

$$\begin{aligned} g_{ij}(x_j, y_j, u_j \,|\, \beta_i, \beta_j, \tau_i, \tau_j) &= 1 + \frac{\rho_{ji}\beta_j\tau_j(x_j - 1)}{N_i} \\ &+ \frac{\rho_{ij}\beta_j\tau_i(y_j - 1)}{N_j} + o(N_1^{-1}) + o(N_2^{-1}) + o(N_1^{-1}N_2^{-1}) \qquad \text{(E.4)} \end{aligned}$$

Integrating over β_i, β_j, τ_j,

$$\begin{aligned} g_{ij}(x_j, y_j, u_j \,|\, \tau_i) &= \int_0^\infty g_{ij}(x_j, y_j, u_j \,|\, \beta_i, \beta_j, \tau_i, \tau_j) \\ &= 1 + \frac{\rho_{ji}R_0^{(j)}(x_j - 1)}{N_i} + \frac{\rho_{ij}\mathbb{E}[\beta_i](y_j - 1)\tau_i}{N_j} \\ &+ o(N_1^{-1}) + o(N_2^{-1}) + o(N_1^{-1}N_2^{-1}) \end{aligned}$$

The PGF for the degree distribution of host node p where $i \in \mathcal{P}_i$ can be calculated as

$$\begin{aligned} G_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u}, |\, \tau_i) &= \mathcal{G}_i\left(\boldsymbol{g_i}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u} \,|\, \tau_i)\right) \qquad \text{(E.5)} \\ &= \prod_{j=1}^{M} \left(g_{ij}(x_j, y_j, u_j \,|\, \tau_i)\right)^{N_j - \delta_{ij}} \end{aligned}$$

Now we apply the limit $N \to \infty, N_j/N \to \eta_j$ in (E.5).

$$\begin{aligned} G_i(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{u} \,|\, \tau_i) = \prod_{j=1}^{M} \exp\Big\{ &\rho_{ji}R_0^{(j)}(x_j - 1)\eta_j/\eta_i \qquad \text{(E.6)} \\ &+ \rho_{ij}\mathbb{E}[\beta_i](y_j - 1)\tau_i\Big\} \qquad \text{(E.7)} \end{aligned}$$

Now the last step is to integrate the conditional distribution with respect to τ_i. Note that in the infinite population limit, undirected host edges disappear from the part of the *percolation network* that excludes the *giant component* [55]. Secondly, the degree distributions of the indegree and the outdegree become independent.

$$G_i(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(\sum_{j=1}^{M} \rho_{ji} R_0^{(j)} \eta_j (x_j - 1)/\eta_i\right) \times \int_0^{\infty} \exp\left(\sum_{j=1}^{M} \rho_{ij} \mathbb{E}[\beta_i](y_j - 1)\tau_i\right) dF(\tau_i) \tag{E.8}$$

We define $G_i^{(in)}(\boldsymbol{x})$ and $G_i^{(out)}(\boldsymbol{y})$ be the PGFs for the indegree and outdegree distributions contained in $G_i(\boldsymbol{x}, \boldsymbol{y})$. Then

$$G_i(\boldsymbol{x}, \boldsymbol{y}) = G_i^{(in)}(\boldsymbol{x}) G_i^{(out)}(\boldsymbol{y}) \tag{E.9}$$

where

$$\begin{aligned} G_i^{(in)}(\boldsymbol{x}) &= \exp\left(\sum_{j=1}^{M} \rho_{ji} R_0^{(j)} \eta_j (x_j - 1)/\eta_i\right) \\ G_i^{(out)}(\boldsymbol{y}) &= \int_0^{\infty} \exp\left(\sum_{j=1}^{M} \rho_{ij} \mathbb{E}[\beta_i](y_j - 1)\tau_i\right) dF(\tau_i) \end{aligned} \tag{E.10}$$

APPENDIX F

Stochastic simulations

In this chapter we describe the methodology for simulating the epidemic models described in the thesis. The models fall in the broader category of continuous time Markov chains and the simulation methods are thus, more general than mathematical epidemiology. We take the example of the simple SIR model with the following rate equations.

$$\begin{aligned}(S, I, R) &\xrightarrow{\beta SI/N} (S-1, I+1, R)\\ (S, I, R) &\xrightarrow{\gamma I} (S, I-1, R+1)\end{aligned} \tag{F.1}$$

As described in the main text, the rates represent probability per unit time of occurrence of a particular reaction, i.e.,

$$\begin{aligned}\mathbb{P}[S(t+dt) = s-1, I(t+dt) = i+1 | S(t) = s, I(t) = i] &= \beta si/N\, dt\\ \mathbb{P}[I(t+dt) = i-1, R(t+dt) = r+1 | I(t) = i, R(t) = r] &= \gamma i\, dt\end{aligned}$$

The process is simulated using Gillespie's direct method [53, 86]. Each transition in the Markov chain requires two uniform random number: one for the time of next transition and the second for the state of next transition. Let the process be in the state (s, i, r) at time t. If the two random numbers are r_1 and r_2, the time of next reaction is $t + \delta t$ where

$$\delta t = -\frac{\log r_1}{\beta si/N + \gamma i} \tag{F.2}$$

The state of the process at time $(t + \delta t)$ is determined by the second random number

$$(S(t+\delta t), I(t+\delta t), R(t+\delta t)) = \begin{cases}(s-1, i+1, r) & \text{if } r_2 < \dfrac{\beta si/N}{\beta si/N + \gamma i}\ ,\\ (s, i-1, r+1) & \text{otherwise.}\end{cases} \tag{F.3}$$

The steps are repeated until the system reaches a fixed point, i.e., a state from which the rate of outgoing transitions are zero. In the SIR process, this happens when I becomes 0.

The method can be easily extended to the case of metapopulation models or the case with constant external forcing. In the first step, the denominator of eq. F.2 would be the sum of rates of all transitions that can occur from the current state. In the second step (eq. F.3), the transition is chosen according to the bin of all possible reactions in which the second random number falls.

References

1. D. Bernoulli, "Essai dune nouvelle analyse de la mortalit cause par la petite vrole," *Mm. Math. Phys. Acad. Roy. Sci., Paris* (1766) 235. 2
2. D. Bernoulli and S. Blower, "An attempt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it," *Reviews in Medical Virology* **14** No. 5, (2004) 275–288. 2
3. K. Dietz and J. Heesterbeek, "Daniel Bernoulli's epidemiological model revisited," *Mathematical Biosciences* **180** No. 1, (2002) 1–21. 2
4. J. Snow, *On the mode of communication of cholera*. John Churchill, New Burlington Street, London, 1855. 2
5. S. Johnson, *The Ghost Map: The story of London's most terrifying epidemic and how it changed science, cities, and the modern world*. Penguin, London, 2006. 3
6. H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review* **42** No. 4, (2000) 599–653. 3
7. W. H. Hamer, *The Milroy lectures on epidemic disease in England: the evidence of variability and of persistency of type*. Bedford Press, London, 1906. 3
8. D. L. Smith, K. E. Battle, S. I. Hay, C. M. Barker, T. W. Scott, and F. E. McKenzie, "Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens," *PLoS Pathogens* **8** No. 4, (2012) e1002588. 3
9. W. Kermack and A. McKendrick, "Contributions to the mathematical theory of epidemics," *Bulletin of Mathematical Biology* **53** No. 1, (1991) 33–55. 3, 18
10. F. Ball, "The threshold behaviour of epidemic models," *Journal of Applied Probability* **20** (1983) 227–241. 10, 59, 100, 111
11. K. B. Athreya and P. E. Ney, *Branching Processes*. Springer-Verlag Berlin Heidelberg, 1972. 10, 11, 18, 53, 61, 62, 95, 132
12. N. T. J. Bailey, *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley-Interscience, 1990. 11, 23, 25, 26, 29, 53, 61, 62, 63, 95, 100, 130
13. F. Brauer, P. Van den Driessche, and J. Wu, *Lecture Notes in Mathematical Epidemiology*. Springer, 2008. 12
14. B. Von Bahr and A. Martin-Löf, "Threshold limit theorems for some epidemic processes," *Advances in Applied Probability* **12** (1980) 319–349. 12
15. T. Antal and P. Krapivsky, "Outbreak size distributions in epidemics with multiple stages," *Journal of Statistical Mechanics: Theory and Experiment* **2012** No. 07, (2012) P07018. 13, 15, 22, 37, 77, 79, 95
16. E. Ben-Naim and P. Krapivsky, "Scaling behavior of threshold epidemics," *The European Physical Journal B* **85** No. 5, (2012) 1–9. 15, 16, 21, 22, 37, 45, 47, 54, 79

17. F. G. Ball, "Dynamic population epidemic models," *Mathematical biosciences* **107** No. 2, (1991) 299–324. 18, 123

18. F. Ball and D. Clancy, "The final size and severity of a generalised stochastic multitype epidemic model," *Adv. Appl. Probab.* **25** (1993) 721–736. 18, 100, 111, 117

19. T. Britton, "Epidemics in heterogeneous communities: estimation of R_0 and secure vaccination coverage," *J. R. Stat. Soc. Series B Stat. Methodol.* **63** No. 4, (2002) 705–715. 18

20. A. Donaldson, J. Gloster, L. Harvey, and D. Deans, "Use of prediction models to forecast and analyse airborne spread during the foot-and-mouth disease outbreaks in brittany, jersey and the isle of wight in 1981," *Veterinary Record* **110** No. 3, (1982) 53–57. 18

21. M. J. Keeling, M. E. Woolhouse, D. J. Shaw, L. Matthews, M. Chase-Topping, D. T. Haydon, S. J. Cornell, J. Kappey, J. Wilesmith, and B. T. Grenfell, "Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape," *Science* **294** No. 5543, (2001) 813–817. 18, 100, 102

22. B. Grenfell, O. Bjørnstad, and J. Kappey, "Travelling waves and spatial hierarchies in measles epidemics," *Nature* **414** No. 6865, (2001) 716–723. 18

23. C. Viboud, O. N. Bjørnstad, D. L. Smith, L. Simonsen, M. A. Miller, and B. T. Grenfell, "Synchrony, waves, and spatial hierarchies in the spread of influenza," *Science* **312** No. 5772, (2006) 447–451. 18

24. H. Andersson and T. Britton, *Stochastic epidemic models and their statistical analysis*, vol. 4. Springer New York, 2000. 21, 53

25. J. P. Sethna, *Statistical mechanics: entropy, order parameters, and complexity*. Oxford University Press Oxford, 2006. 21

26. F. Brauer, P. Van den Driessche, J. Wu, and L. Allen, *Mathematical Epidemiology*. Springer, 2008. 21, 53, 87, 89, 111

27. J. Lloyd-Smith, D. George, K. Pepin, V. Pitzer, J. Pulliam, A. Dobson, P. Hudson, and B. Grenfell, "Epidemic dynamics at the human-animal interface," *Science* **326** No. 5958, (2009) 1362–1367. 22, 23, 54, 56, 96, 97

28. P. Rohani, R. Breban, D. E. Stallknecht, and J. M. Drake, "Environmental transmission of low pathogenicity avian influenza viruses and its implications for pathogen invasion," *Proceedings of the National Academy of Sciences* **106** No. 25, (2009) 10365–10369. 22

29. S. Singh, D. J. Schneider, and C. R. Myers, "Using multitype branching processes to quantify statistics of disease outbreaks in zoonotic epidemics," *Phys. Rev. E* **89**, 032702 (2014). 23, 52

30. A. M. Zubkov, "Life-periods of a branching process with immigration," *Theory of Probability & Its Applications* **17** No. 1, (1972) 174–183. 23, 25, 27

31. S. Ong, "On a class of discrete distributions arising from the

birth-death-with-immigration process," *Metrika* **43** No. 1, (1996) 221–235. 25

32. D. G. Kendall, "Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain," *The Annals of Mathematical Statistics* **24** No. 3 (1953) 338–354. 26

33. D. Shanbhag, "On infinite server queues with batch arrivals," *Journal of Applied Probability* **3** No. 1, (1966) 274–279. 26, 28, 136

34. J. Virtamo, "Queueing theory," *Lecture Notes, Helsinki University of Technology* (2005). 26

35. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in FORTRAN 77: Volume 1, Volume 1 of Fortran Numerical Recipes: The Art of Scientific Computing*, Vol. 1. Cambridge University Press, 1992. 30

36. P. Flajolet and R. Sedgewick, *Analytic combinatorics*. Cambridge University Press, 2009. 31, 142

37. K. Christensen and N. R. Moloney, *Complexity and criticality*, vol. 1. Imperial College Press, 2005. 31

38. J. D. Murray, *Mathematical biology*, vol. 2. springer, 2002. 53

39. M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E* **66** No. 1, (2002) 16128. 53, 75, 87, 97, 109, 110

40. E. Kenah and J. M. Robins, "Second look at the spread of epidemics on networks," *Phys. Rev. E* **76** No. 3, (Sep, 2007) 036113. 53, 75, 97, 104, 105, 108, 109, 110, 111, 146

41. M. E. J. Woolhouse and S. Gowtage-Sequeria, "Host range and emerging and reemerging pathogens," *Emerg. Infect. Dis.* **11** (2005) 1842–1847. 54

42. T. Kuiken *et al.*, "Pathogen surveillance in animals," *Science* **309** No. 5741, (2005) 1680–1681. 54

43. N. D. Wolfe, C. P. Dunavan, and J. Diamond, "Origins of major human infectious diseases," *Nature* **447** No. 7142, (2007) 279–283. 54, 56, 97

44. S. S. Morse *et al.*, "Prediction and prevention of the next pandemic zoonosis," *Lancet* **380** No. 9857, (2012) 1956–1965. 54, 56, 96

45. L. J. S. Allen *et al.*, "Mathematical modeling of viral zoonoses in wildlife," *Nat. Resour. Model.* **25** No. 1, (2012) 5–51. 54

46. S. K. Collinge and C. Ray, *Disease Ecology: Community Structure and Pathogen Dynamics*. Oxford University Press, USA, 2006. 55

47. J. E. Childs, J. A. Richt, and J. S. Mackenzie, "Introduction: conceptualizing and partitioning the emergence process of zoonotic viruses from wildlife to humans," *Curr. Top. Microbiol. Immunol.* (2007) 1–31. 55

48. P. R. Epstein, "Emerging diseases and ecosystem instability: new threats to public health," *Am. J. Public Health* **85** No. 2, (1995) 168–172. 55

49. W. B. Karesh *et al.*, "Ecology of zoonoses: natural and unnatural histories," *Lancet* **380** No. 9857, (2012) 1936–1945. 55, 73, 96, 97

50. T. Antal and P. L. Krapivsky, "Exact solution of a two-type branching process: models of tumor progression," *J. Stat. Mech.* **2011** No. 08, (2011) P08018. 62

51. D. A. Griffiths, "A bivariate birth-death process which approximates to the spread of a disease involving a vector," *J. Appl. Probab.* **9** (1972) 65–75. 63, 95

52. S. Karlin and S. Tavaré, "Linear birth and death processes with killing," *J. Appl. Probab.* **19** No. 3, (1982) 477–487. 63, 67, 95, 140

53. M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, 2008. 65, 91, 149

54. M. E. J. Newman, S. H. Strogatz, and D. J. Watts., "Random graphs with arbitrary degree distributions and their applications," *Phys. Rev. E* **64** No. 2, (2001) 026118. 75

55. E. Kenah and J. M. Robins, "Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing," *J. Theor. Biol.* **249** No. 4, (2007) 706. 87, 105, 106, 110, 111, 146, 148

56. R. Antia, R. R. Regoes, J. C. Koella, and C. T. Bergstrom, "The role of evolution in the emergence of infectious diseases," *Nature* **426** No. 6967, (2003) . 96, 97

57. O. Restif *et al.*, "Model-guided fieldwork: practical guidelines for multidisciplinary research on wildlife ecological and epidemiological dynamics," *Ecol. Lett.* **15** No. 10, (2012) 1083–1094. 97

58. J. L. N. Wood *et al.*, "A framework for the study of zoonotic disease emergence and its drivers: spillover of bat pathogens as a case study," *Philos. Trans. R. Soc. Lond. B Biol. Sci* **367** No. 1604, (2012) 2881–2892. 97

59. J. Hindes, S. Singh, C. R. Myers, and D. J. Schneider, "Epidemic fronts in complex networks with metapopulation structure," *Phys. Rev. E* **88** No. 1, (2013) 012809. 97, 121, 126

60. D. Gatherer, "The 2009 H1N1 influenza outbreak in its historical context," *Journal of Clinical Virology* **45** No. 3, (2009) 174–178. 100

61. L. A. Meyers, B. Pourbohloul, M. E. Newman, D. M. Skowronski, and R. C. Brunham, "Network theory and SARS: predicting outbreak diversity," *Journal of theoretical biology* **232** No. 1, (2005) 71–81. 100

62. V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, and A. Vespignani, "Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions," *PLoS Medicine* **4** No. 1, (2007) e13. 100

63. N. Ferguson, C. Donnelly, and R. Anderson, "The foot-and-mouth epidemic in Great Britain: pattern of spread and impact of interventions," *Science* **292** No. 5519, (2001) 1155. 100, 102

64. S. R. Chowdhury, C. Scoglio, and W. Hsu, "Simulative modeling to control the foot and

mouth disease epidemic," *Procedia Computer Science* **1** No. 1, (2010) 2261–2270. 100

65. S. Riley, "Large-scale spatial-transmission models of infectious disease," *Science* **316** No. 5829, (2007) 1298. 100

66. M. Keeling, L. Danon, M. Vernon, and T. House, "Individual identity and movement networks for disease metapopulations," *Proceedings of the National Academy of Sciences* **107** No. 19, (2010) 8866. 100

67. R. Paysour and W. Fry, "Interplot interference: A model for planning field experiments with aerially disseminated pathogens.," *Phytopathology* **73** No. 7, (1983) 1014–1020. 100

68. L. Madden, G. Hughes, and F. Van den Bosch, *The study of plant disease epidemics*. American Phytopathological Society, 2007. 100

69. M. Keeling and P. Rohani, "Estimating spatial coupling in epidemiological systems: a mechanistic approach," *Ecology Letters* **5** No. 1, (2002) 20–29. 100

70. J. Swinton, "Extinction times and phase transitions for spatially structured closed epidemics," *Bulletin of Mathematical Biology* **60** No. 2, (1998) 215–230.

71. A. Park, S. Gubbins, and C. Gilligan, "Extinction times for closed epidemics: the effects of host spatial structure," *Ecology Letters* **5** No. 6, (2002) 747–755.

72. C. Metcalf, C. Munayco, G. Chowell, B. Grenfell, and O. Bjørnstad, "Rubella metapopulation dynamics and importance of spatial coupling to the risk of congenital rubella syndrome in Peru," *Journal of The Royal Society Interface* (2010) . 100

73. M. Tildesley, T. House, M. Bruhn, R. Curry, M. ONeil, J. Allpress, G. Smith, and M. Keeling, "Impact of spatial clustering on disease transmission and optimal control," *Proceedings of the National Academy of Sciences* **107** No. 3, (2010) 1041. 100

74. N. Ferguson, C. Donnelly, and R. Anderson, "Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain," *Nature* **413** No. 6855, (2001) 542–548. 100

75. A. Allard, P.-A. Noël, L. J. Dubé, and B. Pourbohloul, "Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics," *Phys. Rev. E* **79** No. 3, (2009) 036113. 100, 108, 109, 111, 126

76. G. Scalia-Tomba, "The asymptotic final size distribution of reducible multitype reed-frost processes," *Journal of Mathematical Biology* **23** No. 3, (1986) 381–392. 102, 104, 114, 117

77. L. Mari, E. Bertuzzo, L. Righetto, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "Modelling cholera epidemics: the role of waterways, human mobility and sanitation," *Journal of The Royal Society Interface* **9** No. 67, (2012) 376–388. 102

78. E. Bertuzzo, R. Casagrandi, M. Gatto, I. Rodriguez-Iturbe, and A. Rinaldo, "On spatially explicit models of cholera epidemics," *Journal of the Royal Society Interface* **7** No. 43, (2010) 321–333. 102

79. J. Kephart and S. White, "Directed-graph epidemiological models of computer viruses,"

in *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*, pp. 343–359, IEEE. 1991. 102

80. G. Serazzi and S. Zanero, "Computer virus propagation models," *Performance Tools and Applications to Networked Systems* (2004) 26–50. 102

81. M. E. Newman, "Mixing patterns in networks," *Phys. Rev. E* **67** No. 2, (2003) 026126. 108, 109, 111

82. O. Diekmann, J. Heesterbeek, and M. Roberts, "The construction of next-generation matrices for compartmental epidemic models," *Journal of The Royal Society Interface* **7** No. 47, (2010) 873–885. 109

83. A. Lambert, "Some aspects of discrete branching processes," *Dynamics* **20** (2010) 43. 110

84. H. S. Wilf, *Generatingfunctionology*. Academic Press, San Diego, 1994. 130

85. S. I. Resnick, *Adventures in stochastic processes*. Birkhäuser, Boston, 1992. 130, 135

86. D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry* **81** No. 25, (1977) 2340–2361. 149